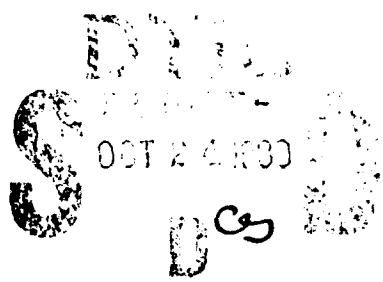
Navy Personnel Research and Development Center

San Diego, CA 92152-6800 TR 89-13 July 1989

(2)

Officer Career Development: Analytic Strategy Recommendations



NPRDC TR 89-13 July 1989

Officer Career Development: Analytic Strategy Recommendations

Lawrence R. James
Christopher K. Hertzog
Georgia Institute of Technology

Reviewed by
Robert F. Morrison

Approved by
John J. Pass
Director, Personnel Systems Department

Released by
B. E. Bacon
Captain, U.S. Navy
Commanding Officer

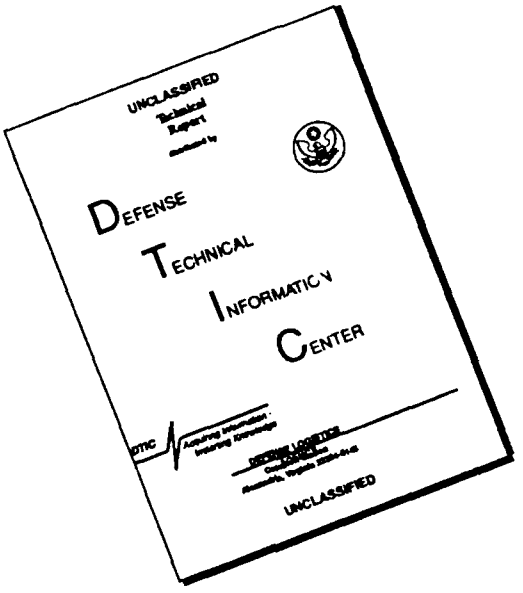
and

James S. McMichael
Technical Director

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-6800

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPRDC TR 89-13	5 MONITORING ORGANIZATION REPORT NUMBER(S) TCN 87-621
6a NAME OF PERFORMING ORGANIZATION School of Psychology Georgia Institute of Technology 6b OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Office
6c ADDRESS (City, State, and ZIP Code) Atlanta, GA 30332	7b. ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING / SPONSORING ORGANIZATION Navy Personnel Research and Development Center 8b OFFICE SYMBOL (If applicable) Code 12	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) San Diego, CA 92152-6800	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO 0602233N; PROJECT NO; TASK NO; WORK UNIT ACCESSION NO 1488 X4B529

11 TITLE (Include Security Classification)
Officer Career Development: Analytic Strategy Recommendations

12 PERSONAL AUTHOR(S)
Lawrence R. James and Christopher K. Hertzog

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM Sep 87 TO Mar 88	1989 July	116

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Analytic strategy, latent variable, time series, cohort analysis, moderator analysis
05	09		

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Strategies are recommended for analyzing information from the data bank developed by the Personnel Distribution and Career Development (PDCD) work unit for the purpose of establishing empirically-based decision guides to assist in the design and implementation of career policy and practice in the U.S. Navy. A set of analytic models is proposed wherein each model addresses an important issue concerning the development of empirically-based decision guides for career development. The statistical assumptions underlying each model are reviewed, as are methods that may be used to reasonably satisfy these assumptions. Estimation techniques and procedures for avoiding common errors in estimation also receive attention.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Robert Morrison	22b TELEPHONE (Include Area Code) (619) 553-9256
	22c. OFFICE SYMBOL Code 12

FOREWORD

This report focuses on a review of the data bank developed with the Personnel Distribution and Career Development (PDCD) work unit data base and the proposed analytical strategy. It describes problems inherent in the data and recommends techniques and strategies to overcome them.

This is the second of two reports completed with TCN 87-621 with Robert F. Morrison as the contracting officer's technical representative. The TCN was conducted within exploratory development (Program Element 0602233N, work unit number 1488WX4B529, Personnel Distribution and Career Development) under the sponsorship of the Chief of Naval Research (ONR 222). This report is the fifteenth published within PDCD and is intended for use in the PDCD work unit.

B. E. BACON
Captain, U.S. Navy
Commanding Officer

JAMES S. McMICHAEL
Technical Director

Prior PDCD Publications

1. Cook, T. M., & Morrison, R. G. (1982, August). Surface warfare junior officer retention: Early career development factors (NPRDC TR 82-59). San Diego: Navy Personnel Research and Development Center.

2. Cook, T. M., & Morrison, R. F. (1983, January). Surface warfare junior officers retention: Background and first sea tour factors as predictors of continuance beyond obligated service (NPRDC TR 83-6). San Diego: Navy Personnel Research and Development Center.

3. Morrison, R. F. (1983, July). Officer career development: Surface warfare officer interviews (NPRDC TN 83-11). San Diego: Navy Personnel Research and Development Center.

4. Morrison, R. F., Martinez, C., & Townsend, F. W. (1984, March). Officer career development: Description of aviation assignment decisions in the antisubmarine warfare (ASW) patrol community (NPRDC TR 84-31). San Diego: Navy Personnel Research and Development Center.

5. University of San Diego (1984, October 23-25). Proceedings: Volume I. Group reports. Tri-service career research workshop. San Diego: Continuing Education, University of San Diego. (Author)

6. Morrison, R. F., & Cook, T. M. (1985, February). Military officer career development and decision making: A multiple-cohort longitudinal analysis of the first 24 years (NPRDC MPL TN 85-4). San Diego: Navy Personnel Research and Development Center, Manpower and Personnel Laboratory.

7. Wilcove, G. L., Bruni, J. R., & Morrison, R. F. (1987, August). Officer career development: Reactions of two unrestricted line communities to detailers (NPRDC TN 87-40). San Diego: Navy Personnel Research and Development Center.

8. Morrison, R. F. (1988, March). Officer career development: URL officers in joint-duty assignments (NPRDC TN 88-26). San Diego: Navy Personnel Research and Development Center.

9. Wilcove, G. L. (Ed.) (1988, August). Officer career development: Problems of three unrestricted line communities (NPRDC TR 88-26). San Diego: Navy Personnel Research and Development Center.

10. Wilcove, G. L. (1988, September). Officer career development: General unrestricted line officer perceptions of the dual-career track (NPRDC TN 88-62). San Diego: Navy Personnel Research and Development Center.

11. Bruni, J. R., & Wilcove. G. W. (1988, October). Officer career development: Preliminary surface warfare officer perceptions of a major career path change (NPRDC TN 89-5). San Diego: Navy Personnel Research and Development Center.

12. Bruce, R. (1989, June). Officer career development: Fleet perceptions of the aviation duty officer program (NPRDC TN 89-25). San Diego: Navy Personnel Research and Development Center.

13. Bruce, R., & Burch, R. (1989, June). Officer career development: Modeling married aviator retention (NPRDC TR 89-11). San Diego: Navy Personnel Research and Development Center.

14. James, L. R., & Hertzog, C. K. (In review). Officer career development: An overview of analytic concerns for the research. San Diego: Navy Personnel Research and Development Center.

SUMMARY

Problem

A large data bank has been developed by the Personnel Distribution and Career Development (PDCD) for the purpose of establishing empirically-based decision guides to assist in the design and implementation of career policy and practice in the U.S. Navy. Data banks of this magnitude often engender special methodological problems during analysis.

Purpose

To recommend analytic strategies that consider not only the special methodological problems that might arise in the analysis of the large data bank but also the need to develop effective and practical models for explaining and forecasting continuance in the Navy, occupational development, and upward mobility in the Navy.

Approach

Analytic strategies are recommended to test causal models for continuance, occupational development, and upward mobility. The strategies involve consideration of (1) the types of analytic models that could be employed to conduct statistical analyses on the data; (2) the conceptual and statistical requirements or assumptions for each analytic model, with accompanying discussion of practical means by which assumptions might be "reasonably satisfied"; (3) actual statistical estimation procedures; and (4) likely specification errors, which refer to problems in estimation and attempts to fit models that occur often in practice.

Emphasis is placed on practical models and designs that provide straightforward means for testing causal models. However, more sophisticated statistical strategies are reviewed in the latter part of the report. Such strategies may be useful for analyses designed for more scientifically oriented audiences.

CONTENTS

LIST OF FIGURES

INTRODUCTION

The objective of this second of two reports is to recommend analytic strategies to test causal models for three key career outcome variables, namely continuance within the Navy, occupational development, and upward mobility within the Navy. This report augments the first report (Report 1), which reviewed the data bank developed by the Personnel Distribution and Career Development (PDCD) work unit in the conduct of research designed to assist in the design and implementation of career policy and practice in the Navy. The first report also considered basic concerns pertaining to analytic strategies for testing career development models. The purpose of this report is to furnish greater breadth and depth in regard to analytic strategies by considering (1) types of analytic models that could be employed to conduct statistical analyses on the data; (2) the conceptual and statistical requirements or assumptions for each analytic model, with accompanying discussion of practical means by which assumptions might be "reasonably satisfied"; (3) actual statistical estimation procedures; and (4) likely specification errors, which refer to problems in estimation and attempts to fit models that occur often in practice.

It is recognized that the PDCD work unit has already devoted considerable time and effort to analytic concerns, including major scaling efforts on the 1982 and 1986 waves of data and development of exploratory models for the outcome variables. It is also recognized that the analytic strategies of paramount importance at the present time are those that will provide the Navy with effective yet practical models for explaining and forecasting continuance, occupational development, and upward mobility. Consequently, we will focus on observed or "manifest" variables designs and both analytic models and statistical strategies that provide straightforward and practical means for testing causal models. We will address the use of more sophisticated analytic models and statistical strategies (e.g., latent variable models) at the conclusion of this report. It is hoped that these discussions will be useful for analyses designed for more scientifically oriented audiences.

This report is presented in four sections that correspond to the natural sequencing of analyses (Skinner, 1978), plus a short summary. Section I addresses scale development. We shall concentrate on potential problems with the use of developed scales in the proposed confirmatory (casual) analyses. Section II pertains to analytic strategies that may be used to test manifest variable causal models within subgroups defined by salient moderators, such as community and career stage. Models, assumptions, statistical techniques, and likely specification errors are considered. Section III is devoted to analytic strategies for comparing the casual models developed in the Section II analyses among two or more subgroups. These are moderator or homogeneity of regression analyses, and models, assumptions, statistical techniques, and likely specification errors are again considered. Section IV is devoted to brief discussions of more sophisticated techniques, including latent variable confirmatory analysis, event-history analysis, and logit analysis. Section V presents a brief summary of key recommendations for future research.

It is noteworthy that this report is designed to present an overview of analytic strategies, with special emphasis on assumptions and potential specification errors. We relied heavily on the published literature from various statistical areas. However, we will be happy to extend and elaborate on special topics in this report, as requested by the PDCD work unit.

SECTION I: SCALE DEVELOPMENT

The decision was made by the PDCD work unit to focus initial research efforts on scales that are common to the two waves of data (i.e., the 1982 wave and the 1986 wave). Inspection of these (common) scales indicates that (1) internal consistency estimates of reliabilities, based on coefficient alpha, tend to be greater than or equal to .75 even though many of the scales (item composites) have only a few items (i.e., three to five items), and (2) the items comprising a particular scale tend, by rational examination, to be assessing a common construct. While much potentially remains to be done regarding tests of the psychometric properties of the data, generally moderate to high reliabilities and scales that make rational sense are good starting points, especially for the practical analyses of primary concern here.

We have one principle concern for these practical analyses. This concern derives from the fact that a large number of manifest causal variables (scales) may be relevant to a particular causal model and thus entered into a confirmatory analysis for that model. In Report 1, we noted that use of a potentially large number of manifest scales in a confirmatory analysis increases the probability of multicollinearity (Gordon, 1968). Products of multicollinearity include large standard errors for ordinary least-squares (OLS) coefficients (regression weights), which spuriously detracts from findings of significant relations, and instability in the OLS estimates themselves (cf. Johnston, 1984). We recommended use of latent variable designs as a possible solution to the potential multicollinearity problem. However, given the decision to proceed initially with manifest variable designs, alternatives are needed. We suggest the following procedures.

1. Correlations among causal variables entering into a particular equation for OLS analyses or an overall model for LISREL analyses need to be examined. A "very high correlation" (e.g., > .75) suggests the possibility of an ensuing multicollinearity condition in the regression/LISREL analysis.

2. Examination of bivariate correlations is often not sufficient to identify potential multicollinearity conditions because no one bivariate correlation is very high. However, one or more causal variables may be linearly dependent on some subset of the remaining causal variables, which does create a multicollinearity condition. Checks for linear dependence may be made by regressing each causal variable in a causal system (e.g., causal or structural equation) on the other causal variables in that system (i.e., each of K causal variables is regressed on the remaining K-1 causal variables). If the squared multiple correlation (i.e., R^2 or SMC) for a particular variable is high, then this variable may be linearly dependent on the other variables in the system and inclusion of this variable in analyses may create a multicollinearity condition. (Common factor analysis programs often furnish the R^2s of interest here inasmuch as R^2s--SMCs--are often used as initial estimates of communalities.

3. Results of confirmatory analyses should be checked carefully for indications of multicollinearity, or "near multicollinearity" (Johnston, 1984, p. 245). Very large standard errors for regression coefficients, estimated regression coefficients that change with small changes in the data (e.g., random addition or deletion of a small number of cases, a large R^2 with few significant regression coefficients, and a pattern of bivariate correlations are indicative of multicollinearity and near multicollinearity problems.

4. There exist numerous remedies to the (near) multicollinearity problem (see Johnston, 1984, pp. 250-259). The most direct and practical remedies are:

a. Delete some causal variables from a set of highly correlated causal variables. For example, if one has three causal variables that intercorrelate .90, then drop two of the variables.

b. Form a composite of highly correlated causal variables. This alternative accomplishes some of the same objectives as a latent variable approach, given that the manifest variables to be combined are measures of the same construct. We recommend that only indicators of the same construct be combined. Theory, substantive content of variables, bivariate correlations, and perhaps a factor analysis could be employed to ascertain whether variables are measures of the same construct. We might also note that we prefer this alternative to the deletion of variables, a key reason being that reliabilities of the variables used in analyses are likely to be enhanced by forming composites.

c. Use block-recursive forms of analyses (cf. Namboodiri, Carter, & Blalock, 1975, pp. 526-530). Block recursive analysis is similar to regression analyses based on sets of independent variables (cf. Cohen & Cohen, 1983, Chapt. 4), and is often applied in complex designs. Sets of theoretically related variables are identified and grouped into blocks of variables (e.g., environmental, career counseling, motivation, affect, etc.). A causal model is then constructed for, in this case, a single dependent or "endogenous" variable (e.g., career intent), but the causal mechanisms are represented by blocks of variables rather than by single variables. Analysis then proceeds by introducing one block of variables at a time into an OLS equation--that is, a hierarchical regression analysis (see Cohen & Cohen, 1983, Chapt. 4). For each block of variables, only the change in the R^2 is interpreted (i.e., the degree to which introduction of this set of variables enhanced prediction). No attempt is made to interpret the regression weights for individual variables (within blocks) because of the likelihood of multicollinearity.

In sum, we suggest the judicious use of alternative "b" (forming combinations) when combinations of variables are clearly indicated, followed by the use of block recursive models if multicollinearity still appears to be a problem, which is quite possible in complex designs involving many causal variables. Later, in Section IV, we shall address additional scaling issues. Of special concern is the use of latent variable models to compare factor structures (measurement models) over subgroups defined by key moderator variables such as career stage and cohort.

SECTION II. ANALYTIC STRATEGIES FOR INITIAL TESTS OF CAUSAL MODELS

The general model of career development proposed by Morrison and Cook (1985) and reviewed in Report 1 suggests that it is unlikely that a single causal model will suffice to explain all continuance decisions (or all decisions pertaining to either occupational development or upward mobility). Rather, a series of moderators likely bound or limit the generalizability of a particular causal model to an identifiable subset of the data (i.e., a subgroup). Three potentially salient sources of moderation are: (1) community (SWO, AWO URL(G)) as well as subcommunities within communities (e.g., AWO-P and AWO-NFO); (2) career stage, which refers to key career choice points (Morrison, 1983) and was illustrated in terms of "social cohorts" (Morrison & Cook, 1985) in Report 1 (see p. 11), and (3) generational differences, which refers to basic differences among the members of different cohorts. Note that career stage refers to a form of sequential moderation

wherein causal models for career decisions differ for the same individuals over time (Ghiselli, 1956; James, Joe, & Irons, 1982; James & Tetrick, 1984), whereas generational differences refers to variations in causal models for different groups of individuals defined by year of commissioning.

It is anticipated that the PDCD work unit will combine career development theory, knowledge of Navy practices, and empirical data to define meaningful subgroups for analyses. (If possible, please note our recommendation in Report 1 to avoid clustering by empirical similarity using profile analytic techniques.) We devote Section II of this report to analytic strategies for initial tests of causal models within the subgroups so defined by the PDCD unit. Section III addresses comparisons of models among subgroups--that is, moderator analyses. Statistical recommendations are made in Section II that will prepare the data and initial results for the moderator analyses proposed in Section III.

Analytic Models

We begin by briefly reviewing the types of manifest variable analytic models that potentially could be applied to the Navy career development data to answer salient, practical problems. As presented in Report 1, these analytic models include:

1. Cross-sectional models (Figure 1a): The key to these models is that all data were collected at approximately the same time for a particular individual. An example is a model developed for the 1982 wave (or the 1986 wave) data to explain career intent for officers in the SWO community who have been in the Navy for 18 to 30 months.

2. Longitudinal model (Figure 1b): As applied to this study, a longitudinal model is typically one in which the data on causal variables are collected cross-sectionally by questionnaire, but data on the key endogenous (criterion, dependent) variable is collected at a later date. An obvious example is a combination of the cross-sectional model illustrated above with data on continuance (retention) collected on a longitudinal basis. Additional illustrations of this form of model are presented in Figures 2 and 3 of Report 1.

3. Nonlagged, cross-sectional time series (Figure 1c): As shown in Figure 4 of Report 1 and as discussed on pages 15 and 16 of that report, this form of analytic model requires that repeated measurements by taken on multiple individuals at two or more points in time and (a) all causal effects take place within specified time intervals and (b) there are no lagged causal effects from one time interval to the next time interval (cf. Nerlove, 1971; Hannan & Young, 1977; Johnston, 1984). It is unlikely that this model will receive much attention in the career development research because of the number of hypothesized lagged effects in the Morrison and Cook (1985) career development model.

4. Lagged cross-sectional time series (Figure 1d.): The lagged form of cross-sectional time series is again based on repeated measures from multiple individuals over time. Here, however, variables measured at one point in time (e.g., 1982) are causes of variables measured at another point in time (e.g., 1986). When an endogenous variable such as career intent is viewed as a cause of itself over time (see Figure 1d and pages 15 and 17 in Report 1), then the model takes the form of a "lagged endogenous variable, cross-sectional time series" (cf. James & Singh, 1978; Johnston, 1984; Ostrom, 1978). Unfortunately, with but two waves of measurement, the model is not a complete lagged endogenous variable, cross-sectional time series because a third wave of data is needed to test key hypotheses and to effect what are likely the most appropriate statistical analyses. Nevertheless, it is expected that this analytic model will be useful in the practically oriented analyses of primary concern here. Consequently, we will devote

considerable attention to this model. Note also the opportunity to add longitudinally measured endogenous variables (e.g., continuance (y_4)) to the design.

FIGURE 1a. CROSS-SECTIONAL MODEL (1982 QUESTIONNAIRE)

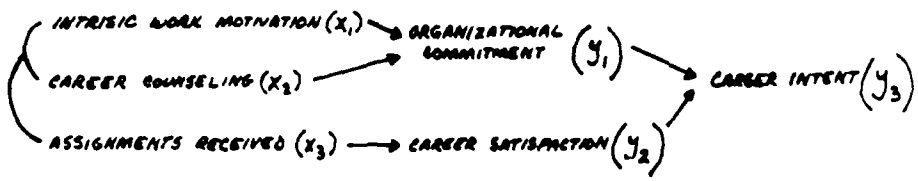


FIGURE 1b. LONGITUDINAL MODEL
(X_1 THROUGH Y_3 SAME AS FIG. 1a, Y_4 MEASURED LONGITUDINALLY)

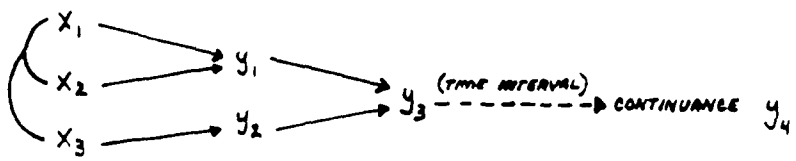


FIGURE 1c. NONLAGGED CROSS-SECTIONAL TIME-SERIES

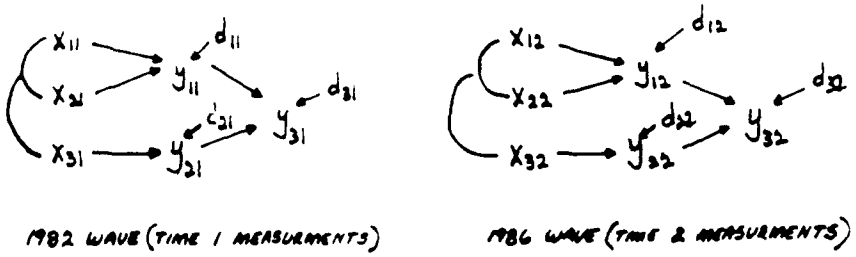


FIGURE 1d. LAGGED ENDOGENOUS VARIABLE

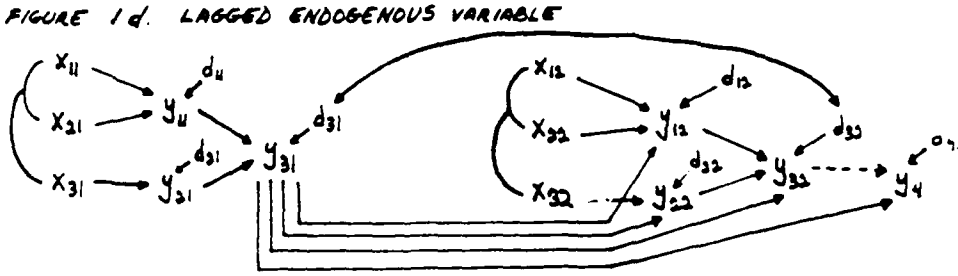


Figure 1. Potential analytic models for confirmatory analyses.

In sum, we have four analytic models that, while not exhaustive of all possible analytic models, will be the key models used to test causal hypotheses within subgroups defined by salient moderators. As noted, the nonlagged cross-sectional time series has limited applicability and thus is not considered further in this section of the report. Each of the three remaining analytic models could be employed to test salient hypotheses for each of the three criteria. Cross-sectional analyses could be conducted for the 1982 and/or the 1986 waves of data (within subgroups) for endogenous variables represented by "decisions" collected by means of questionnaires (see document entitled Outcome Variables: Career Decisions and Actions). Longitudinal analyses could be conducted for the

1982 and/or 1986 waves for each of the three key endogenous variables because each such variable has a longitudinal component (represented by "actions" in the document noted above). Finally, the lagged cross-sectional time series design is applicable for individuals who have both 1982 and 1986 questionnaire data.

Conceptual and Statistical Requirements for Analytic Models

We shall focus here on general conditions that are required to subject a theoretical model to confirmatory analysis, as discussed by James, Mulaik and Brett (1982), and on general statistical assumptions required of manifest level confirmatory analyses, with additional attention to specific assumptions required for longitudinal models and lagged cross-sectional time series. Statistical assumptions that are associated with specific estimation techniques are addressed later in discussions of these techniques.

The seven conditions pertaining to the appropriateness of theoretical models for confirmatory analysis presented by James et al. (1982) are reproduced in Figure 2. The extensive theoretical development and modeling that preceded development of the multiple questionnaires suggests reasonable satisfaction of Conditions 1 and 2 (cf. Morrison & Cook, 1985). (By reasonable satisfaction, we refer to what is scientifically acceptable even though imperfect.) Condition 6--specification of boundaries--pertains primarily to the moderator analyses (nonadditivity) that are the subject of the Section III of this report. Condition 4--specification of causal direction--is, like Conditions 1 and 2, already reasonably satisfied inasmuch as the career development models to be tested in the initial analyses are "recursive" (i.e., all causal relations are unidirectional). Later, in more scientifically oriented analyses, the PDCD work unit may wish to consider tests of selected nonrecursive relations inasmuch as the Morrison and Cook (1985) general model of career development presumes a number of dynamic relations (see James & Jones 1980; James & Singh, 1978; James & Tetrick, 1984 for illustrated uses of nonrecursive models in psychology). Nonlinearities in some causal relations, an issue included in Condition 6, might also be considered in these later analyses.

This leaves us with Conditions 3, 5, and 7 as pertinent to the case at hand. We begin with Condition 7, which states that structural (causal) models should be stable. Stability is indicated by invariance of values of structural parameters over specified time intervals, which technically is referred to as "stationarity." Appropriate lengths of time intervals vary with variables and models, but the general idea is that a time interval should be of sufficient length to allow for scientific inferences and generalizations. On the other hand, there is no assumption that the model or structural parameters are set in concrete. That is, change in the parameters is allowed over different time intervals, such as different career stages. Indeed, stability of structural parameters across different time intervals (career stages) is an empirically testable question if data are available.

Stationarity is testable using both the cross-sectional and longitudinal models. Indeed, these will be salient concerns in the moderator tests discussed in Section III. Stationarity of the lagged cross-sectional time series, or lagged CSTS, cannot be tested until a third wave of data are collected.

A point related to both stability and Condition 3 (specification of causal order) is that the values on the variables in the structural equations should have reached a state of approximate constancy before data were collected. This assumption is referred to as the "equilibrium-type condition" (cf. Namboodiri et al., 1975) and is predicted on the logic that confirmatory analysis is designed to ascertain if a hypothesized causal model(s) could have generated a particular set of data. That is, the causal processes are assumed to have

Condition 1:	Formal statement of theory in terms of a structural (causal) model. Development of a structural model that specifies variables, causal connections among variables, and functional relations and equations that relate each effect to all of its relevant causes.
Condition 2:	Theoretical rationale for causal hypotheses. Use of theory to propose how causes produce effects by introduction of mediating mechanisms to help to explain nonobvious covariation among variables causal connections among complex variables.
Condition 3:	Specification of causal order. Hypothesized order in which variables occur naturally in a causal model, given an equilibrium-type condition for cross-sectional designs and specified causal intervals, stationarity, and an equilibrium-type condition for time series designs.
Condition 4:	Specification of causal direction. Hypothesized direction of causation for each causal connection in a structural model. The direction may be asymmetric, denoting a recursive causal relation, or reciprocal, denoting a nonrecursive causal relation.
Condition 5:	Self-contained causal equations. The causal equation for each effect (endogenous variable) in a structural model contains all the relevant causes of that effect, which is indicated by lack of covariation between the explicitly measured causes in an equation and the disturbance term of that equation.
Condition 6:	Specification of boundaries. Given linearity in parameters and variables, the causal equations are additive within the populations (e.g., subjects and environments) to which inferences are to be made.
Condition 7:	Stability of structural model. The values of structural (causal) parameters are invariant (stationary) over specified time intervals, and the values on variables representing events are in an equilibrium-type condition.

Figure 2. Conditions pertaining to appropriateness of theoretical models for confirmatory analysis. (Adopted from James, Mulaik, & Brett, 1982, Figure 2.6, pp. 56-57).

already taken place and their effects to have worked their way through the system so that the system is in a state of temporary equilibrium. The confirmatory analyses designed to determine if a model(s) has a good fit with the data is thus essentially inquiring whether this model(s) could have generated these data. To answer this question requires first that

the causal processes have occurred and that the values on the variables have reached a state of temporary constancy--an equilibrium-type condition.

Estimators of certain types of stability, such as test-retest reliability, provide at least indirect tests of the equilibrium-type condition (indirect because they require only a correlational form of reliability). Most important, however, is the concern that individuals should have been in the Navy and in their positions for a sufficient period of time to be able to respond meaningfully to the questionnaires. Specifically, whatever causal influence is indicated by a questionnaire item should have already occurred. It is suggested that the PDCD work unit consider carefully whether all members of the data set had been in position for sufficient periods of time for causal effects to have stabilized. A final point in this regard is the use of the equilibrium-type condition to establish a causal order in the cross-sectional and longitudinal designs. As discussed in greater detail in James et al. (1982, pp. 51-54), length of causal intervals, or equilibrium times, may be used to establish causal orders and to avoid the infinite regress implied by many open system, dynamic models.

Otherwise, the specification of causal orders required by Condition 3 is largely provided by theory (cf. Morrison & Cook, 1985). And, it is possible and legitimate to propose several alternative causal orders a priori, conditional on having good theoretical reasons for each alternative ordering, and to conduct tests to ascertain which one of the orderings best fits the data (Billings & Wroten, 1978; see James & Tetrick, 1984 for an example). In fact, proposing multiple, alternative, theoretically-based models for the same set of data, and contrasting these a priori models in terms of fit with the data, is a highly recommended approach to confirmatory analysis (cf. James et al., 1982, Chapt. 3). On the other hand, one should not explore different causal orders with the same set of data in order to find the causal order that has the best fit with the data (Duncan, 1975). This is never a legitimate exercise and, if attempted, one that is almost surely to be heavily criticized. (A middle ground is changing causal orders as part of a specification analysis. Such changes should be few and theoretically based. Of course, if theoretically based then they might have been a priori, thereby perhaps obviating the need for a specification search.)

A final point regarding causal order pertains to the lagged CSTS, where ordering for some aspects of the causal model are determined by time of measurement (i.e., 1982 or 1986). While use of CSTS reduces at least some possible ambiguities in causal ordering (e.g., events in 1986 could not have caused events in 1982), there is a price to be paid in the use of CSTS, or with any form of time series or panel-type design. This price is the requirement that the times of measurement (measurement intervals, such as the interval between the 1982 and 1986 waves of measurement) "must correspond closely to the true causal intervals in a time-series design (Kenny, 1979)" (James et al., 1982, p. 37). It is recommended, therefore, that the PDCD work unit give special attention to a theoretical justification for the causal interval for any lagged effect (e.g., a causal connection between a 1982 variable and a 1986 variable). Moreover, the causal intervals will vary among individuals in the longitudinal designs (e.g., all sample members took the questionnaire in 1982, but those who did not continue in the Navy left at different times). Length of time between questionnaire administration and continuance action should thus be considered in terms of theoretical implications and perhaps treated as a variable.

The final condition, and perhaps the most salient one, is Condition 5, which requires that causal equations be self-contained. Statistically, self-containment requires that no covariation occur between causal variables included explicately in a structural equation and the (theoretical) disturbance terms of that equation (James, 1980; James et al., 1982;

Johnston, 1984). Note that this assumption is based on the theoretical disturbance in a structural model and structural equations and not on the residual or error terms used to estimate disturbances by statistical analyses. A less statistically oriented approach to this assumption is to require that all relevant causes of an (or each) endogenous variable are included in the structural equation for that endogenous variable (James et al., 1982). A relevant cause is a causal variable that (1) has at least a moderate, direct effect on the endogenous variable (2) is stable, (3) is related to at least one other causal variable in the structural equation, and (4) is not linearly dependent on the other causes in the causal equation.

The basic idea of self-containment, or its obverse, the unmeasured variables problem, is that no key causal variable is left out of a causal equation. But, of course, this is unavoidable because current scientific knowledge regarding most endogenous variables, including career decisions and actions, is incomplete and thus all relevant causes cannot be considered to be known. Reasonable satisfaction of the self-containment condition requires attempts be made to include known relevant causes in structural equations (James et al., 1982). A set of decision criteria for establishing reasonable satisfaction of Condition 5 is presented in James (1980). Since these criteria are rather extensive, they are not reproduced here. However, the James (1980) article is included as Appendix A.

General Statistical Requirements for Confirmatory Analysis

The following overview of statistical requirements was obtained from many sources, principal among these were Bentler and Chou (1987), Duncan (1975), Hayduk (1987), Heise (1975), Johnston (1984), Joreskog and Sorbom (1986), Kenny (1979), Long (1983a, 1983b), Namboodiri et al. (1975), and Ostrom (1978). Salient statistical requirements that must be satisfied by all of the three analytic models (cross-sectional, longitudinal, lagged CSTS) are presented below. While lengthy, the list is not exhaustive of every possible requirement, and several important assumptions are addressed in the discussion of estimation techniques. Moreover, we have not differentiated between assumptions required for estimation of parameters and assumptions required only for interpretation of parameters and statistical inference, the logic being that one usually wishes to interpret what one has estimated.

The equation below is presented to assist in the discussion of assumptions.

$$Y = A + B_1X_1 + B_2X_2 + d \qquad (1)$$

where Y is the endogenous variable, X_1 and X_2 ($X_j, j = 1,2$) are causal variables, B_1 and B_2 ($B_j, j = 1,2$) are structural parameters for the X_j in raw-score or deviation-score form (if Y and the X_j are in standardized form, then the B_j would be path coefficients), A is the intercept, and d is the disturbance.

1. Within subgroups defined by salient moderators, relations represented by the structural parameters B_1 and B_2 are linear and additive. The issue of nonadditivity or moderation (or interaction) is, as noted earlier, the subject of Section III, and thus this issue is not discussed here. Linearity (in the variables) refers to the form of functional relationship linking the endogenous variable to the causal variables. For example, a simple bivariate relation is linear if it can be represented by a straight line having the form $Y = A + bX$, plus error in stochastic models. Nonlinearity in the variables is often addressed by polynomial regression equations (cf. Cohen & Cohen, 1983), where one or more continuous causal variables is (are) raised to powers (typically squared) to represent nonlinear functions such as U or inverted-U shaped relations in the bivariate case.

2. If X_1 and/or X_2 is a continuous variable, then the scale of measurement is at least interval. As noted by James et al., (1982), an essentially interval scale reasonably satisfies this assumption (see Boyle, 1970).

3. The causal variables are perfectly reliable (if variables, including Y are in standardized form, then all variables in the model are assumed to be perfectly reliable). James et al. (1982) suggest that "high" reliabilities reasonably satisfy this assumption, but note the lack of consensus of a criterion for what constitutes "high." Nevertheless, the generally high coefficient alphas for the majority of (questionnaire) variables included in this study suggests that problems due to random measurement errors (e.g., attenuation) are unlikely to be substantial. Use of latent variable models in future efforts should reduce the problem even further.

4. The X variables are not linearly dependent. We have already discussed this issue in regard to its role in multicollinearity.

5. The disturbances have a multivariate normal distribution, where each disturbance has a mean of zero, and the variances of the disturbances are equal. These are standard assumptions for statistical techniques such as OLS (ordinary least squares), and involve well known assumptions such as normal distributions of the Ys within arrays and homoscedasticity.

6. X_1 and X_2 are nonstochastic or fixed variables. Confirmatory analysis is often based on an OLS (ordinary least squares) "fixed" regression model wherein the Y and d are random variables and the X_j are fixed variables. This fixed variable regression model is perhaps better suited to experimental designs where investigators determine discrete values for each X_j and then randomly sample subjects into these values. Nevertheless, popular texts such as Cohen and Cohen (1983) are based on the fixed variable regression model and this model is often used to analyze data where at least some of the X_j are clearly random variables.

Relaxing this assumption and allowing the X_j to be stochastic or random variables is necessary given that many if not most of the causal variables in the career development models are random variables. This is easily accomplished if one is willing to assume that (a) conditional on each X (i.e., X_1 and X_2), the disturbances are normally and independently distributed with means equal to zero and variances equal, and X_1 and X_2 are unrelated to d, which is the self-containment condition discussed earlier in regard to Condition 5. With these assumptions, the use of traditional OLS procedures will furnish meaning estimators and significance tests, especially in large samples (see Cramer & Appelbaum, 1978; Johnston, 1984, Chapt. 7).

7. Absence of nonrandom measurement errors. A nonrandom measurement error is a systematic source of basis that, if present, reduces the accuracy with which a manifest variable represents an underlying construct or latent variable (Namboodiri et al., 1975). As reviewed in James et al. (1982, p. 58), nonrandom measurement errors involve (a) aggregation and disaggregation biases, (b) ceiling and floor effects in measurement scales, (c) classification errors resulting from poor scaling of manifest variables (e.g., reducing a reliable continuum to a dichotomy), (d) method variance resulting from the fact that two or more manifest variables share a common measurement procedure and thus are influenced by common response sets/response biases, and (e) serially correlated errors of measurement that result from use of the same measurement scale(s) in two or more waves of data collection.

The career development data, like almost any set of field data collected in part by questionnaires, is likely subject to several of these types of errors. Aggregation bias is not a problem as long as individual level data are analyzed with individuals as the unit of analysis. (Unit and/or macro level variables may be added to these analyses using techniques discussed by James, Demaree, and Hater, 1980--see Appendix B). Aggregation of individual level data and analyses of such aggregate data should proceed only after careful consideration of issues pertaining to cross-level inference (see Pedhazur, 1982, pp. 526-547 for a brief and cogent review of the issues).

The investigators should already be aware of ceiling/floor effects and classification errors that may exist in the data, given their prior scaling efforts. Thus, we proceed to the question of method variance. Tests for method variance are often based on application of confirmatory factor analysis (CFA) to various operationalizations of the multitrait-multimethod matrix (cf. Widaman, 1985; Schmitt & Stults, 1986). Such tests, however, require that each construct (latent variable) be measured by using at least two different methods. Generally, this is not an option with the career development data.

A less desirable but applicable alternative often employed by James and colleagues (e.g., James & Jones, 1980) is designed to test whether a pervasive method factor has biased questionnaire data. To illustrate the use of this procedure, suppose we have three constructs, labelled A, B, and C. All constructs are measured by the same procedure (e.g., a questionnaire). Theory may suggest a high correlation between A and B. Suppose a high correlation is obtained. Suppose further that a critic argues that this high correlation is primarily a product of method variance (i.e., a pervasive method factor created a spurious correlation between A and B). A test of the critic's argument is provided by introducing variable C, where (1) C is measured in the same manner as A and B, (2) C is subject to the same response sets/styles as A and B (e.g., acquiescence), (3) C has psychometric characteristics that are similar to A and B, and (4) C theoretically has low relationships with A and B. Now, with these conditions, high correlations between C and both A and B implies a pervasive method factor. However, low correlations between C and both A and B suggests the absence of a pervasive method factor and thus the high correlation between A and B cannot be totally spurious. Other levels of correlation between C and both A and B suggest varying levels of partial spuriousness engendered by a pervasive method factor.

The final concern in regard to nonrandom measurement errors is correlated measurement errors. Such correlations can be easily checked in future analyses on the CSTS models that employ latent variable designs (cf. Joreskog & Sorbom, 1986 and Section IV).

Additional assumptions for cross-sectional time-series. In addition to the above, use of (lagged) cross-sectional time series requires reasonable satisfaction of the following assumptions. We present these assumptions using the lagged CSTS model in Figure 3 as a guide. In Figure 3, d_0 and y_0 represent theoretical measurements that are included to denote that the time t data cannot be analyzed by themselves without creating a serious unmeasured variables problem. (Please note the implications of this point for Figure 1d.) Time t is analogous to the 1982 wave data, whereas Time t + 1 represents the 1986 wave data. Time t = 2 refers to a future wave of data collection. The structural equations for Figure 3 are (variables are assumed to be in deviation form):

$$y_{t+1} = B_t y_t + B_{1t+1} x_{1t+1} + B_{2t+1} x_{2t+1} + d_{t+1} \quad (2)$$

$$y_{t+2} = B_{t+1} y_{t+1} + B_{1t+2} x_{1t+2} + B_{2t+2} x_{2t+2} + d_{t+2} \quad (3)$$

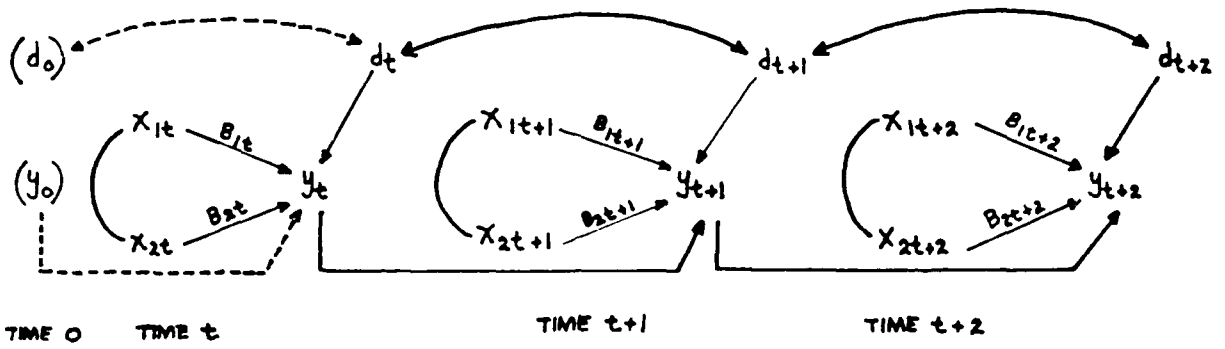


Figure 3. An example of a lagged, cross-sectional, time-series model.

The equations state that y (e.g., continuance intention) is a function of y at a prior time and contemporaneously assessed exogenous variables x_1 and x_2. Note that no equation exists for y_t, which again is due to unavailability of y_o.

The assumptions unique to this lagged CSTS are:

8. Times of measurement correspond to causal intervals, which has been discussed.

9. The model is stationary, which, based on prior discussion, would be indicated by $B_t = B_{t+1}$, $B_{1t+1} = B_{1t+2}$, and $B_{2t+1} = B_{2t+2}$ in Equations 3 and 4.

10. Disturbances are nonautoregressive, which means that no covariation exists between d_{t+1} (Equation 2) and d_{t+2} (Equation 3).

Without the time t + 2 measurements, there is no way to test for stationarity (a test for stationarity has been provided by James & Tetrick, 1984). Moreover, given the likelihood of unmeasured causal variables, it is probable that Assumption 10 above will be violated (see James & Singh, 1978, Figure 8). This is because lack of autocorrelation between d_{t+1} and d_{t+2} (or between d_t and d_{t+1}, see Figure 3) presumes that the disturbances are composed of random shocks (or is a white noise series--cf. Johnston, 1984, p. 371). If this is the case, then the structural parameters in Equation 2--the only estimable equation given two waves of measurement--can be estimated directly with no further ado.

However, consider now that unmeasured relevant causes reside in the disturbance terms (cf. James, 1980), and it is these unmeasured relevant causes that are, in part, responsible for the autocorrelation of the disturbances (the curved arrows between the d's in the model). Inasmuch as no field model is self-contained, it follows that the disturbances will be autocorrelated. Straightforward estimation is no longer possible. Various complex forms of instrumental variables, generalized least squares, or maximum likelihood (Johnston, 1984, Chapt. 9) are required. This is a moot point, however, because, without a third wave of data, most of these complex forms of analyses cannot be implemented. Consequently, the investigators will have to decide whether the key, known relevant variables are included explicitly in their lagged CSTS equations having the form

of Equation 2. If this is believed to be the case, then they may proceed to estimate parameters. These estimates will be both biased and inconsistent, and significance tests will be more powerful than they should be (see Ostrom, 1978). Nevertheless, these problems should not be of great magnitude, or at least of a sufficient magnitude to preclude analyses.

Statistical Estimation Procedures

Most of the career development models involve continuous variables up to the point of the final, endogenous action or outcome variable. The action or outcome variable may be continuous, as in the case of the upward mobility measures, dichotomous, as in the case of continuance, or nonordered and discrete (i.e., qualitative), which applies to occupational development. In the last case--the final analyses involving the occupational development action variables--a multiple discriminant analysis (MDA) will likely be in order. For the continuous upward mobility variable and the dichotomous continuance variable, we suggest the use of OLS or, preferably, maximum likelihood (ML) analyses. Later, in more sophisticated analyses, the dichotomous continuance variable can be subjected to such things as event history analysis (cf. Allison, 1984) and/or logit analysis (cf. Berry & Lewis-Beck, 1986).

With the exception of the use of MDA to complete the analyses on occupational development, the statistical estimation issue boils down to whether one is going to use single equation estimation techniques (OLS) versus full-information estimation techniques (LISREL).

To address this issue, consider the structural equations for the cross-sectional model presented in Figure 1a (variables are in deviation form):

$$y_1 = b_{y} \times x_1 + b_{y} \times x_2 + d_y \quad (4)$$

$$y_2 = b_{y} \times x_3 + d_y \quad (5)$$

$$y_3 = b_{y} \times y_1 + b_{y} \times y_2 + d_y \quad (6)$$

A single equation estimator such as OLS could be used to estimate the structural parameters in each of the three equations. The term "single equation estimator" denotes that a separate OLS analysis is conducted for each equation and thus the estimating process for one equation is independent of the estimating process for another equation. Consequently, specification errors that engender bias or inconsistency in one equation do not spread over and affect the bias or consistency of estimates in another equation (unless the second equation is subject to the same specification errors base don its own lack of merits).

In contrast, a full-information estimator, such as the full-information ML procedures used in LISREL (Joreskog & Sorbom, 1986), would estimate all the structural parameters in Equations 4 through 6 simultaneously. While more efficient, the full-information techniques suffer the problem that specification error in one equation can spread over and affect estimates in a different equation (cf. MacCallum, 1986). On the other hand, a salient benefit of full-information techniques is opportunity to test the overall fit of the model to the data. In this regard, we strongly recommend reading of Wheaton (1987). Moreover, use of the full-information techniques in LISREL will assist substantially in proceeding to the moderator analyses discussed in Section III.

In sum, either OLS or LISREL may be used for estimation purposes. We recommend LISREL, which generally means the use of full-information ML to analyze manifest variable structural models. Checks may be made to compare the LISREL estimates to OLS estimates. If the estimates differ, then a potential culprit is the spreading of specification errors by the full-information technique. If this appears likely, then the OLS estimates would be preferable.

The statistical assumptions required to employ OLS are as discussed, with one addition. The addition is that each equation must be identified. Identification refers to the question of whether sufficient information is available to obtain unique mathematical estimates of structural parameters (cf. James et al., 1982). Recursive equations based on manifest variables are generally identified and thus we will not pursue this issue here.

For full-information maximum likelihood (FIML), one must assume that the Xs and Ys are distributed multivariate normal. In addition, identification must be established for each parameter (cf. Long, 1983b). The identification issue should again not be a problem.

There are many additional issues that will occur during statistical analyses. We prefer to deal with these issues on an interactive basis with the investigators as they arise. On the other hand, we do wish to reiterate several points raised in Report 1 that are germane to estimation. These points are (1) use of hold-out samples for cross-validation purposes (cf. Cudeck & Brown, 1983); (2) avoidance of the use of change scores; (3) use of nonstandardized data in analyses, especially the lagged CSTS; and (4) development of comparison and generalization samples (see p. 22, Report 1) for lagged versus nonlagged analyses and for analyses based on selected samples (e.g., selection of equal numbers of stayers and leavers).

Statistical Specification Errors

Section II is concluded with a listing of errors that occur often in practice during estimation and model fitting (cf. Bentler & Chou, 1987; Billings & Wroten, 1978). We focus on issues that were not addressed in prior discussions of conditions for causal modeling, statistical assumptions, and estimating techniques. We recommend Bentler and Chou (1987) for elaboration on points 4 through 7.

1. **Sample sizes that are too small for stable statistical results.** This concern may arise in the career development study as a result of subgrouping, which is to say that one or more of the subgroups defined by salient moderators is too small. While no clear-cut criterion exists for defining "small" (there are many heuristics, however), our experience suggests that attempts be made to keep sample sizes above 200.

2. **Restriction of variance on the criterion.** Low variance on a criterion (endogenous variable), which is often associated with a skewed distribution, is associated with problems in trying to predict/explain occurrences of the criterion, especially if data are standardized (e.g., path analysis). This problem may be a result of naturally occurring events, such as low base rates, or induced, such as restriction of range due to incidental selection. Remedies include the use of correction equations, the use of unstandardized data, and the use of samples selected to remove base-rate problems.

3. **Presence of outliers in the data.** Outliers may or may not affect various aspects of analyses. An article by Stevens (1984) has a good review of procedures for detecting outliers.

4. **Use of distribution free methods on small samples.** Bentler and Chou (1987) recommend that unweighted least squares (URL--a full-information method in LISREL) be used only when n > 200.

5. **Failure to use multiple test and fit criteria to evaluate a causal model.** It has become apparent to many authors that a model should be subjected to multiple tests and evaluated with multiple fit indices (see Wheaton, 1987).

6. **Use of significance tests on standardized data in LISREL.** The chi-square significance testing procedures are designed for unstandardized data only (Bentler & Chou, 1987).

7. **Failure of estimation procedure to converge in LISREL.** Bentler and Chou (1987) suggest that failure to converge may be due to (a) a nonlinear model that is treated as if linear, (b) a very poor initial model, (c) poor start values for parameters, (d) unreasonable equality constraints, and (e) unidentified, initial parameters.

SECTION III. MODERATOR ANALYSES

A major issue for the project staff is the likelihood that causes of major variables related to Navy career decisions may differ across subcommunities, different time periods, officer ranks, cohorts, and other variables. Our discussion with the principal investigator and his staff have made it clear that detection of such moderator variables is a crucial and primary goal of the research. Detection of moderator variable influences are obviously necessary for accurate projection of future trends and complete understanding of officer career development processes.

Analysis of moderator influences can proceed using either the ordinary least squares (OLS) or maximum likelihood (ML) approaches to manifest variable designs. There are two chief design features to consider in designing the analysis. The first is whether the moderator variable is by nature a categorical or continuous variable. The second is whether the hypothesized locus of moderation requires testing of moderator influences between independent samples or, alternatively, tests of moderator influences between different equations within the same sample (primarily, in testing differences in regression equations in lagged panel data).

Categorical Moderator Variables

Independent Groups Analyses

Analysis of moderator variables is simple and straightforward if the moderator variable is a naturally occurring categorical variable, such as officer cohort. Here we assume that (1) the variables involved in the regression equations are equivalently measured across levels of the categorical moderator, (2) there is sufficient sample size at each level of the moderator variable to permit meaningful statistical analysis in each subgroup, and (3) the analysis is to be done with metric regression coefficients. Condition (1) would be violated in many instances in comparisons of subcommunities, where different variables are measured and where, in some instances, variables have a materially different interpretation in, say, aviators than in surface warriors. In such cases, formal moderator analysis is not warranted. Analysis would proceed independently for each subcommunity, but the regression equations would not be analytically tested for equivalence across subcommunities. Condition (3) is crucial. In general, one does not

wish to estimate the moderator effects in groups where separate standardization of variables has occurred. Separate standardization reduces the likelihood of cross-validation of regression coefficients, in general. In the case of moderator analysis, it is inappropriate to test for interaction if different transformations have been applied to different groups. Separately standardizing the groups is one such case. Calculation of different transformations can introduce or obscure interaction effects. Thus, the analysis cannot be done by analyzing correlation matrices for each of the groups. This would generally not be done in the multiple regression approach, in which group membership is treated as a variable and data from the entire sample is analyzed. Separate standardizations could easily be requested when using LISREL to do the simultaneous equations approach. This is inappropriate, and the analysis should be conducted on covariance matrices of the manifest variables. It is perfectly acceptable to standardize the variables for construction of composites, but this standardization must be done on the pooled data prior to segregation into groups for moderator analysis.

The moderator analysis proceeds in two different ways, depending upon whether separate regression equation or simultaneous estimation approaches are employed (see Section II).

Separate Regression Equations. Moderator analysis proceeds by using product variables and hierarchical regression techniques (Cohen & Cohen, 1983; Pedhazur, 1982). If all exogenous and endogenous variables in the regression equations are continuous variables (excepting the moderator(s)), then product variables are created by multiplying the continuous variables by a set of coded vectors representing category membership. A categorical moderator with m levels will require m-1 coded vectors, unless more restrictive a priori hypotheses about moderation are to be entertained. We generally favor orthogonal coding for representation of moderators, although other coding approaches can be used. A separate three-stage hierarchical regression is then performed for each regression equation from the overall model. In stage 1, all independent variables for the equation are entered and an R^2 and regression coefficients are estimated. In stage 2, the coded vectors representing the moderator groups are added to the equation. In stage 3, the product variables are added. The significance of the increment to R^2 from stage 2 to stage 3 is the critical test of whether there is interaction between the moderator variable and the other variables entered in stage 1. The appropriate statistical test is the traditional F-test for the increment to R^2. It can be requested directly in some statistical packages (e.g., SPSSX Regression, which we recommend generally for hierarchical regression because of ease of interpretation of output). As discussed rather nicely by Pedhazur (1982), significant interactions, if present, mandate calculation and comparison of separate regression equations for each group (categorical level of the moderator; see below). In the absence of moderation, the common (pooled) regression coefficients estimated at stage 1 in the analysis may be used as estimates of effects.

The analysis can become quite cumbersome if (1) multiple moderator variables must be considered simultaneously, and (2) if there are many levels of each moderator. Our assessment of the data set is that this is not generally the case, and that the use of hierarchical regression approaches will prove satisfactory in many cases.

Simultaneous Equation Analysis. If full information maximum likelihood (FIML) approaches have been used, then an alternative approach to moderator analysis can be executed by using LISREL VI or VII (Joreskog & Sorbom, 1984). Although there are other excellent FIML programs for structural regression models (such as Bentler's EQS program), LISREL is the only program currently containing the option to analyze regression

equations simultaneously in multiple groups. Henceforth, we shall discuss the simultaneous equation approach presuming use of LISREL, but the work group should be aware that BMDP, distributors of EQS, have issued pre-release publicity about a version 3.0 of EQS that apparently will handle multiple groups analyses. Thus, the moderator analysis approaches described below may, in the near future, be possible in EQS.

The multiple groups approach is the basis for testing moderator variables in LISREL. One begins by cutting the categorical variable into mutually exclusive groups and specifying the causal model in each group. Then the appropriate test of moderation is whether the unstandardized regression coefficients are equal across the multiple groups. This hypothesis is easily tested in LISREL by testing a model in which the regression coefficients are constrained equal.

The formal statistical test of moderation requires two separate models. In the first model, one simply runs the regression analysis simultaneously in each group. Assuming that the model is overidentified, then this analysis produces a likelihood ratio (LR) X^2 test of the goodness of fit of the regression model to the data. The LR test is, in essence, the sum of the LR X^2 across all the moderator groups. Then one runs a second model that specifies the exact same regression model but also specifies that the regression coefficients are equal in the multiple groups. This second model is said to be nested in the first model, because it has the same basic specification but imposes the additional constraint that the coefficients, which ar free to vary between groups in the first model, are required (constrained to be equal in the second model) (e.g., Dwyer, 1983; Hayduk, 1987; Joreskog & Sorbom, 1979; Long, 1983b). The LISREL program estimates the common regression coefficients but also produces a new LR X^2. Because the two models are nested, the difference in LR X^2s is a formal test of the null hypothesis that all regression coefficients are equal in all groups. The LR X^2 must be greater for the more restricted model with equality constraints. However, if there is no moderation, so that the regression coefficients are truly equal across the subpopulations, then the two LR X^2 tests will be approximately equal, except for sampling error (i.e., the difference in the two X^2s will be approximately equal to the difference in degrees of freedom (df)). Thus, the test for moderation is to calculate the difference in LR X^2, calculate the difference in df (which should equal the number of regression coefficients times m-1, where m is the number of groups), and evaluate the LR X^2 difference against a critical value of the X^2 distribution. It must be emphasized that this test of moderation is a multivariate significance test of moderation across all regression equations.

Use of the LISREL approach to moderator variables is actually quite efficient. One does not need to generate coded vectors, product variables, and test hierarchical increments to R^2. One gets a single, overall test of moderation across all equations. This efficiency in the statistical test may be a curse rather than a blessing, however, if it is expected in advance that the moderator variable affects relatively few of the overall number of regression equations. In that case, the Type II error rate of the overall LR X^2 test for the few equations that are truly different across groups will be higher than in single equation approaches. On the other hand, the simultaneous approach provides better control of the Type I error rate across all equations than the separate equation approach. This is not merely a function of the fact that an overall LR test is computed (as opposed to separate F-tests for each equation). Given that the same independent variables are usually present in multiple regression equations (which is the case if endogenous variables are specified to have both direct and indirect effects), the separate regression equations will not be statistically independent. Although the regression coefficients in a single equation are independent across multiple, independent groups, the regression coefficients will have nonzero covariances of estimate across different equations owning to shared

independent variables. The LISREL approach takes these covariances of estimate into account in calculating the overall LR X^2 test. The separate regressions approach does not.

It is possible to get a separate LR X^2 test for each equation that is an exact logical analog of the F-test for each equation in the separate equation approach. Again, the LR X^2 test is superior in that the covariances of estimate are still used in calculating the statistical test of fit. This is done by imposing the equality constraints on only one equation at a time, and then calculating the difference in X^2 against the model with no equality constraints on any equation. Moreover, it is possible to specify a priori that there will be moderation on a subset of equations and to test a LISREL model specifying moderation only on these equations. In the case of mixtures of equality constraints and no equality constraints across equations, the specification of the model is somewhat more complicated. One has to specify the equality constraints parameter by parameter, but this is by no means a major obstacle.

Frankly, the use of LISREL to test moderator influences is relatively new, and has not been widely discussed in the literature. Hayduk (1987), for example, does treat the issue of "stacked models" but devotes more space to the issue of simultaneous models for means and covariance structures than to the implications of testing equivalence of structural coefficients across multiple groups. When moderator analysis in LISREL is discussed, it is usually at the level of latent variable rather than manifest variable designs. There is insufficient simulation data to evaluate differences between the two methods (separate equation hierarchical regression and LISREL). Our limited experience suggests that the two methods produce quite similar results in recursive models. Additionally, there are advantages for simultaneous approaches such as LISREL above and beyond efficiency. They are also appropriate for nonrecursive models and for models with correlated regression residuals, both of which are poorly handled by single equation procedures (with or without moderator variables). The specification of the equality test in LISREL is also very simple. By far the most difficult problem is specifying the base model, although this is also relatively straightforward in LISREL analysis with manifest variables. We can provide a sample LISREL specification of the equality test upon request.

Post-hoc Comparisons. The significance tests for either the separate equation approach or the simultaneous equation approach are tests of what might be termed "omnibus" null hypotheses (i.e., no moderator effects on any independent variable in the equation). If the null hypothesis is rejected, the alternative hypothesis is that not all regression coefficients are equal across all groups. At that point, a second analysis is required. The purpose of this analysis is to determine (1) which moderator groups differ on (2) which regression coefficients. Without prior theory, this approach entails post-hoc comparisons of regression coefficients across groups. The logic of the post-hoc analysis is exactly the same as the more familiar post-hoc tests of means in ANOVA.

We will describe a general approach for testing the moderator effects across independent groups. It should be noted that this approach does not take into account the full covariance matrix of the regression estimates in calculating specific error terms for post-hoc comparisons. This approach is easily applied for post-hoc detection of moderator effects with both the single equation and simultaneous equation approaches. The analysis proceeds from the parameter estimates from the regression equations of each group. In the single equation approach, one must first calculate separate regression equations in each moderator group (which had not been done prior to the detection of significant interaction). In the LISREL approach, one uses the estimated regression coefficients and

standard errors from the model that imposed no equality constraints on the parameters. Here we assume that one has available (1) regression coefficients for all equations for all groups, and (2) standard errors of estimate for all coefficients. It is possible to compare any pair of regression coefficients by use of the following formula:

$$\underline{t} = b_1 - b_2 / se_{comp},$$

where $\underline{t}$ is a $\underline{t}$-test, b_1 and b_2 are regression coefficients, and s_{comp} is the standard error for the comparison. The formula can be generalized to any linear combination of regression coefficients (see below). The standard error for the comparison, when the regression coefficients are derived from independent samples is

$$se_{comp} = (\text{var est } (b_1) + \text{var est } (b_2))^{1,2},$$

where var est is the variance of estimate for the regression coefficient. Most regression packages report both the b-weight and the standard error of the b-weight (which is the square root of the variance of estimate), so the comparison and the standard error of the comparison are easily calculated. LISREL reports the ML parameter estimates and, upon user request, their asymptotic standard errors. Parenthetically, it should be noted that LISREL's t-values test the null hypothesis that the population parameter is equal to zero. They are not the $\underline{t}$-test of the difference in regression coefficients over moderator groups described above.

The problem, of course, is that there is a very large number of such comparisons that can be made as the number of equations, independent variables per equation, and moderator group levels increase. Practically, it can become quite tedious to calculate the $\underline{t}$-test for all comparisons. From a statistical inference perspective, the more important problem is protection of the Type I error rate across multiple comparisons. Corrections for all possible pair-wise combinations of regression coefficients would probably be too conservative (have too high a Type II error rate). In our opinion, the best approach is to employ a Bonferroni correction on the critical value for $\underline{t}$ used in evaluating the comparisons. The Bonferroni approach maximizes statistical power while controlling the Type I error rate (see Ramsey, 1982). With the Bonferroni approach, one adjusts the critical value of $\underline{t}$ according to the actual number of comparisons to be entertained. Maximum power is achieved by sequential adjustment of the critical value, but this is tedious in practice.

This approach should provide a reasonable degree of protection of Type I error rate while minimizing Type II errors. It assumes that the covariances of estimates for all parameters is zero. As discussed above, this assumption is violated when the same independent variables appear in multiple equations. The principal assumption of importance is undoubtedly independence across levels of the moderator variable, which is satisfied by the independent groups analysis. Nevertheless, it is possible in the simultaneous equation approach to generalize the post-hoc analysis by computing linear contrasts across the vector of regression coefficients and then creating an appropriate standard error for the contrasts by pre- and post-multiplying the covariance matrix of the estimates by the vector of contrasts. At this point, it is not known what the inflation of the Type I error rate is under these conditions, although it seems plausible that the degree of inflation is minimal. This problem has not, to our knowledge, been simulated in the statistical literature. Our recommendation is to proceed with the simpler post-hoc comparisons under independence assumptions (particularly if the single equation approach is used).

This recommendation is driven by pragmatic constraints. It is too time-consuming to calculate the asymptotically exact standard errors of the contrasts by hand by using the covariance matrix of the estimates. Indeed, it will be tedious to compute the pair-wise comparisons by hand, even when using the simultaneous Bonferroni adjustment to the critical value of the test statistic and only employing the standard errors of estimate. In principle, it would be a straightforward programming task to create a program to generate post-hoc statistics on the regression coefficients, incorporating sequential Bonferroni adjustments, use of the entire covariance matrix of estimates to generate standard errors, and options for setting (and perhaps changing) the desired experiment-wise Type I error rate. Employment of the covariance matrix of estimates is most easily and efficiently done using the LISREL program and the simultaneous equation approach. LISREL can, upon user request, output both the regression parameter matrix and the covariance matrix of the estimates. From these matrices, the appropriate t statistics can be calculated directly in matrix form. It would therefore be possible to program the asymptotically unbiased post-hoc tests using the entire covariance matrix of the estimates. Indeed, our past experience with algorithms of this type is that most of the programming overhead involves constructing the input matrices and the formatting of the output statistics rather than the statistical algorithms. Thus the difference in programming time for the comparisons using the entire covariance matrix of the estimates would differ trivially from comparisons using only the standard errors of estimate. The PDCD unit may wish to devote some programmer time to development of this program.

Within-groups Moderator Analysis

The within-groups moderator analysis is required whenever equations are to be compared across variables measured on the same persons. One example is where equations are to be compared across multiple work settings (e.g., prediction of satisfaction in multiple settings from the same background variables). A unique but potentially important case in the PDCD data sets would be the special case of time of measurement as a categorical moderator variable. The concept is actually inherent in the lagged endogenous models discussed above. For example, one might wish to know if the within-occasion predictors for intent to remain in the Navy have changed from 1982 to 1986 for the same officer cohort. Provided that the independent variables are scaled in the same way (although if not, the test can still be performed after judicious rescaling of the predictors) at both occasions, it is possible to test for stationarity in the prediction equations by doing a repeated measures test of the equality of the regression equations over time. This sort of approach is conceptually related to but distinct from the lagged endogenous causal models discussed in Sections I and II, but may be important for forecasting purposes. The basic logic would also apply, however, to testing for equality of effects in the lagged endogenous causal models.

The crucial issue here, from a statistical perspective, is that the regression equations are correlated because they are calculated on the same observational units. Thus, any test of the equality of the regression coefficients from these related equations must take the covariances of estimate from the regression equations into account (James & Tetrick, 1984).

Separate Regression Equations. James and his colleagues (James, Joe, & Irons, 1982; James & Tetrick, 1984) have outlined a procedure by which testing of related regression equations may be accomplished. The procedure involves (1) calculation of regression coefficients for the separate equations, and (2) use of asymptotic theory to derive an appropriate estimate of the covariance matrix of the estimates for the related equations. These computations are relatively tedious, but manageable. Using the matrix of

estimated regression coefficients from the related equations and the approximate covariance matrix of these estimates, one can derive F-tests for the comparisons of subsets of regression coefficients. The approach is quite general and powerful, and handles the case of planned comparisons. Copies of the relevant papers have already been provided to the work group.

Simultaneous Equation Approach. It is possible to use the LISREL program to test the equality of the related regression equations. Unlike the independent samples case described above, the data are treated as a single group. However, the full system of regression equations is specified, with a separate equation for each moderator variable. So, in the case of time of measurement, one would simultaneously specify the equations for the 1982 and 1986 waves in the same model. The same logic with respect to statistical inference applies. One first computes the model for all equations, and (assuming an over identified model) obtains the LR X^2 statistic and the parameter estimates. One then runs a second model in which equality constraints are imposed on those coefficients that are hypothesized to vary across levels of the within-groups moderator variable. The difference in LR X^2 tests is a test of the null hypothesis that the regression coefficients are equal over time.

One advantage of the LISREL approach is that it is possible to combine the subgroup and related-equations moderator analysis into a single model, when appropriate.

Continuous Moderator Variables

Analysis of continuous moderator variables poses additional problems. It is relatively simple to use hierarchical regression techniques to test for differential impact on continuous moderator variables, provided that one assumes the nature of moderation is linear and continuous across the range of the moderator and other independent variables (James, 1987). (In fact, this assumption is often unreasonable). The test procedure is identical to that with categorical moderator variables, in that product variables are formed by multiplying the moderator variable(s) and the independent variable(s). From there, the same three-stage process is employed. Stage 3 involves adding the product variables to the equation and testing the increment to R^2. If the product terms do not increase prediction, then one can consider using the equations from stage 1 or stage 2, depending upon (1) the logical status of the moderator as a predictor variable in the system of equations and (2) the significance of the regression coefficients involving the candidate moderator variable. In practice, continuous variable interactions of this type are not usually moderator variable analyses per se, but rather tests of additivity of causal influence across endogenous variables. So in all likelihood, the variables were already in the system of equations and would be kept in the final equations.

The real sticky wicket is what to do if continuous variable interactions are detected. The two chief regression texts that review these issues (Cohen & Cohen, 1983; Pedhazur, 1982) differ quite dramatically on what to do in such circumstances. Pedhazur eschews the practice generally, for reasons we do not find compelling. Cohen and Cohen (1983) suggests nested substitution of equations so as to be able to graph the nature of the interaction. This is descriptively informative but not necessarily sufficient. Rules for calculating direct and indirect effects in structural equation models in the presence of interaction have been discussed perfunctorily in the sociological literature. Conceptually, the problem is analogous to the more familiar issue in ANOVA: how does one interpret main effects (linear effects of variables) in the presence of interaction (continuous variable moderation)? The answer in regression, as in ANOVA, is that the simple linear effect of x on y, in the presence of interaction involving x and z, is descriptively

meaningful (in essence, consistent direction but variable magnitude of relationship of x and y over the entire range of the moderator, z) but uninterpretable from a causal point of view. Small wonder most investigators choose not to even look for interactive effects (assuming they do not exist) or, alternatively, convert moderators to categorical variables to assist in ease of computation and interpretation. As pointed out by Cohen and Cohen (1983), this approach throws away information if the independent variables are both continuous and the interaction is linear in both variables. Of course, given discontinuous interaction effects, the grouping approach may be superior, provided that the proper cutoffs for assigning groups is known a priori, stumbled upon by chance, or detected by interpretation of scatter plots, regression residuals, and other techniques (James, 1987).

The real problem is how to introduce continuous variable moderation into simultaneous equation approaches. Here the simplicity of the testing procedures in single equation approaches is appealing. The problem is that introduction of the product variables into the regression equations causes a specification error in terms of the hypothesis of uncorrelated errors in equations. It also introduces correlations between regression coefficients and disturbances, which must be modeled explicitly if the regression coefficients and associated standard errors are to be unbiased (by specification error). Kenny and Judd (1984) have discussed this issue in latent variable modeling (see also Hayduk, 1987). The only method for handling this type of analysis is to use covariance structure models like COSAN that can impose nonlinear constraints on parameter estimates. We suspect--and hope--that much more will be known about this problem in a few years. For now, we suggest that tests of continuous interaction be entertained on theoretical grounds, and investigated using hierarchical regression, if needed. We do not believe enough is known about the introduction of product variables into structural equation models to justify staff effort to learn the nuances of nonlinear constraint specification and how to use the COSAN program (which makes LISREL look like BASIC).

SECTION IV. FUTURE DIRECTIONS

The intent of this section is to furnish recommendations for future research that has a scientific emphasis. We will be brief because we wish only to highlight possible avenues for work group consideration. On the other hand, we are prepared to work with staff at this time on these methods if they are considered desirable for immediate emphasis and evaluation.

Use of Categorical Endogenous Variables

Some of the endogenous variables in the data set are true categorical variables. We have discussed using true categorical variables as moderators, but this mainly applies when the categorical variables divide individuals into mutually exclusive groups. When outcome (criterion) measures are categories, the project team may prefer to predict the criterion rather than test for moderation by it. For example, a crucial problem is predicting the retirement decision (stay in the Navy, opt for retirement, opt for retraining, etc.). Knowing which variables provide prediction of the outcome categories is different than asking whether other variables differ in relationship according to outcome category. The situation is made more complex when the prediction equation is actually nested in a multiple equation structural model.

Experts differ on whether one can introduce categorical endogenous variables into linear structural models of the kind we have been discussing. We admire the courage of

Bentler and Chou (1987), who categorically state, without much supporting argumentation or data, that this approach is fully acceptable provided that the marginal frequencies are not excessively disproportional (e.g., an 80% to 20% split) and becomes more acceptable as the number of categories increases. There are other alternatives that can be considered. One is the use of logit regression to predict the categorical dependent variable. We would generally recommend this approach if (1) the single equation approach has been used and (2) even if not, if the categorical variable is a final outcome, like retirement decision. More elegant analysis for categorical variables include techniques like latent class analysis, including the Grizzle/Kock/Landis approach for GLS estimation of effects, and event history analysis. The latter is akin to logit regression but takes into account, and indeed models explicitly, the time course of shifts in category group membership. Allison (1982, 1984) provides a useful introduction to this set of techniques. We do not recommend this as a general approach for the research team at this time, given the relatively limited time remaining to analyze the data set.

Latent Variable Structural Equation Models

Although we have recommended manifest variable designs, given the time constraints on the project, it would be preferable to conduct analyses using the full LISREL approach, particularly on the lagged endogenous variable models. We wish to discuss briefly the benefits of doing the full latent variable models.

The chief benefit of the latent variable models is that the structural regression coefficients are disattenuated for measurement error. It is frequently astonishing to observe the degree of impact measurement error can have in structural equation models when single indicators have moderate reliabilities. As we point out in Section I, the reliabilities reported for the candidate variables are encouraging, and composite variables are usually more reliable than their individual constituent variables. Nevertheless, it would be desirable to estimate effects without contamination of measurement error.

Another related benefit of structural equation models with latent variables is the opportunity to test directly assumptions of equivalence of the measurement model in lagged designs. We often assume, by fiat, that composite variables measure the same construct in equivalent ways across time (or across groups). The chief advantage of longitudinal measurement models is the ability to test equivalence in the measurement model using the type of LR X^2 tests described above, but where the tests are test of constraints on the regression coefficients of observed variables on latent variables (rather than tests on the structural regression coefficients themselves). Hertzog (1987) has reviewed some studies that have employed this approach in examining adult intellectual development and measurement properties of mood state variables (see also Hertzog & Nesselroade, 1987; Hertzog & Schaie, 1986, 1988 in Appendices C through E for detailed examples).

Another advantage of latent variable models is the commensurate increase in the validity of the regression coefficients. Provided that there is minimal sharing of method variance, the structural regression estimates from latent variable models are more likely to represent construct relationships than systematic measurement (method) variance.

Another useful application of structural modeling is in the domain of confirmatory factor analysis itself. Although manifest variable designs with composites can be appropriate, they are more fully justified if it can be shown that the indicators do indeed factor as hypothesized (perhaps implicitly) by the compositing scheme. Thus, it is possible to do confirmatory factor analysis to justify compositing variables, and then use the composites to test the continuous interaction hypotheses using hierarchical regression.

The composites can also, in such conditions, be formed by use of factor score estimation procedures rather than simple unit-weighting of z-scores. Hultsch, Hertzog, and Dixon (1984) used this approach to examine age X intelligence interaction effects in predicting text memory in adults (see Appendix F for details).

SECTION V. SUMMARY

We have outlined a series of research design and analysis options for staff to consider. In a report like this, it is difficult to specify exactly what an appropriate structural model would look like. This is best done in direct design consultation with the contractors on a specific research problem, bringing theory about measurement and latent variable relationships to bear in the design phase.

Our general recommendation has been for the work group to proceed immediately with manifest variable regression analysis that has predictive utility, is more easily summarized and communicated to higher levels in the Navy, is scientifically defensible, and can be accomplished in relatively short order. This decision is driven in large part by pragmatic considerations. An alternative is for the work group to decide to take additional time and to concentrate on some of the latent variable techniques described in the last section of the report. It is important to note that, should the work group decide to pursue latent variable structural equations analysis, then it is advisable to consider a roughly two-stage process (Anderson & Gerbing, 1988): (1) development of the measurement model for all exogenous and endogenous latent variables with confirmatory factor analysis, followed by (2) incorporation of the structural regression model into the previously developed measurement model. This approach has two advantages. First, one can be confident that the structural model is not contaminated by specification errors in the measurement model. Usually, it is the structural coefficients that are of primary interest, and one does not want spread of specification error from the measurement model in an FIML approach to bias structural regression coefficients. Second, it is possible to treat the full structural model as a more restricted, nested model from the measurement model, and to then calculate a difference in X^2 statistic that separates lack of fit in the structural model from the overall fit of the model, combining lack of fit in both structural and measurement submodels. This approach provides a more accurate assessment of the viability of the structural model.

The implication of the foregoing is that, if the PDCD work group decides to proceed with latent variable rather than manifest variable modeling, then the immediate strategy should be to begin work on the confirmatory factor analysis of multiple indicators for latent variables rather than computation of composites. In either case, we look forward to working with the group in adapting the general principles described here to specific analyses.

REFERENCES

Allison, P. D. (1982). Maximum likelihood estimation of linear models when data are mission (Unpublished manuscript).

Allison, P. D. (1984). Event history analysis: Regression for longitudinal event data (Sage University Paper Series/Number 07-046). Beverly Hills and London: Sage.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411-423.

Bentler, P. M., & Chow, C. P. (1987). Practical issues in structural modeling. Sociological Methods and Research, 16, 78-117.

Berry, W. D., & Lewis-Beck, M. S. (Eds.). (1986). New tools for social scientists: Advances and applications in research methods. Beverly Hills and London: Sage.

Billings, R. S., & Wroten, S. P. (1978). Use of path analysis in industrial/organizational psychology: Criticisms and suggestions. Journal of Applied Psychology 63(6), 677-688.

Boyle, R. P. (1970). Path analysis and ordinal data. American Journal of Sociology, 75, 461-480.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cramer, E. M., & Appelbaum, M. I. (1978). The validity of polynominal regression in the random regression model. Review of Educational Research, 48, 511-515.

Cudeck, R., & Brown, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-167.

Duncan, O. D. (1975). Introduction to structural equation models. New York: Academic Press.

Dwyer, J. H. (1983). Statistical models for the social and behavioral sciences. New York: Oxford University Press.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. Journal of Applied Psychology, 40, 374-377.

Gordon, R. (1968). Issues in multiple regression. American Journal of Sociology, 73, 592-616.

Hannan, M. T., & Young, A. A. (1977). Estimation in panel models: Results on pooling cross-sections land time series. Social Methods, 52-53.

Hayduk, L. A. (1987). Structural equation modeling with LISREL. Baltimore: Johns Hopkins University Press.

Heise, D. R. (1975). Causal analysis. New York: Wiley.

Hertzog, C. (1987). Applications of structural equation models in gerontological research. In K. W. Schaie (Ed.), Annual review of gerontology and geriatrics, 7, 265-293. New York: Springer.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. Child Development, 58, 93-109.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. Psychology and Aging, 1, 159-171.

Hertzog, C., & Schaie, K. W. (1988). Stability and change in adult intelligence: 2. Simultaneous analysis of longitudinal means and covariance structures. Psychology and Aging, 3, 122-130.

Hultsch, D. F., Hertzog, C., & Dixon, R. A. (1984). Text processing in adulthood: The role of intellectual abilities. Developmental Psychology, 20, 1193-1209.

James, L. A. (1987). Continuous vs. discontinuous moderation: A case for segmentation (Unpublished doctoral dissertation). Atlanta: Georgia Institute of Technology.

James, L. R. (1980). The unmeasured variable problem in path analysis. Journal of Applied Psychology, 65, 415-421.

James, L. R., Demaree, R. G., & Hater, J. J. (1980). A statistical rationale for relating situational variables and individual differences. Organizational Behavior and Human Performance, 25, 354-364.

James, L. R., Joe, G. W., & Irons, D. M. (1982). A multivariate test for sequential moderation. Educational Psychological Measurement, 42, 951-960.

James, L. R., & Jones, A. P. (1980). Perceived job characteristics and job satisfaction: An examination of reciprocal causation. Personnel Psychology, 33, 97-135.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills: Sage.

James, L. R., & Singh, K. (1978). An introduction to the logic, assumptions, and basic analytic procedures of two-stage least squares. Psychological Bulletin, 85, 1104-1122.

James, L. R., & Tetrick, L. E. (1984). A multivariate test for homogeneity of regression weights for correlated data. Educational and Psychological Measurement, 44, 769-780.

Johnston, J. J. (1984). Econometric methods. New York: McGraw-Hill.

Joreskog, K. G., & Sorbom, D. (1979). Advances in factor analysis and structural equation models. Cambridge, MA: Abt Associates.

Joreskog, K. G., & Sorbom, D. (1986). LISREL VI analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. Morresville, IN: Scientific Software, Inc.

Kenny, D. A. (1979). Correlation and causality. New York: Wiley.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. Psychological Bulletin, 96, 201-210.

Long, J. S. (1983a). Confirmatory factor analysis: A preface to LISREL. Beverly Hills: Sage.

Long, J. S. (1983b). Covariance structure models: An introduction to LISREL (Sage University paper Series/Number 07-034). Beverly Hills and London: Sage.

MacCallum, R. (1986). Specification searches in covariance structure modeling. Psychological Bulletin, 100, 107-120.

Morrison, R. F. (1983). Officer career development: Surface warfare officer interviews (NPRDC Tech. Note 83-11). San Diego: Navy Personnel Research and Development Center.

Morrison, R. F., & Cook, T. M. (1985). Military officer career development and decision making: A multiple-cohort longitudinal analysis of the first 24 years (NPRDC MPL Tech. Note 85-4). San Diego: Navy Personnel Research and Development Center.

Namboodiri, N. K., Carter, L. R., & Blalock, H. M., Jr. (1975). Applied multivariate analysis and experimental designs. New York: McGraw-Hill.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross sections. Econometrica, 39, 359-382.

Ostrom, C. W. (1978). Time series analysis: Regression techniques. Beverly Hills: Sage.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd Ed.), New York: Holtz, Rinehart, and Winston.

Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on p variables. Journal of Educational Statistics, 7, 139-156.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

Skinner, H. A. (1978). The art of exploring predictor-criterion relationships. Psychological Bulletin, 85, 327-337.

Stevens, J. F. (1984). Outliers and influential data points in regression analysis. Psychological Bulletin, 95, 334-344.

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. Sociological Methods and Research, 16, 118-154.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

APPENDICES A THROUGH F

SUPPLEMENTARY READINGS

APPENDIX A

THE UNMEASURED VARIABLES PROBLEM IN PATH ANALYSIS

By

Lawrence R. James
Institute of Behavioral Research
Texas Christian University

Journal of Applied Psychology
1980, Vol. 65, No. 4, 415-421

The Unmeasured Variables Problem in Path Analysis

Lawrence R. James
Institute of Behavioral Research, Texas Christian University

The unmeasured variables problem has not received adequate attention in applications of path analysis. The ramifications of inadequate attention to this problem are addressed in respect to correlations between causal variables and the errors of causal equations and the resulting bias in solutions of path coefficients. The discussion recognizes that obviation of the unmeasured variables problem is an unrealistic objective. Consequently, logic is provided in the form of decision steps to help investigators ascertain whether the influence of unmeasured variables that can be expected in any particular analysis is of sufficient seriousness to preclude the use of path analysis.

In their review of path analysis studies, Billings and Wroten (1978) concluded that many biased estimates of path coefficients had been reported in the industrial organizational literature. A primary reason cited was that relevant causal variables had not been included in the causal systems investigated. This unfortunate practice is generally referred to as the unmeasured (or omitted) variables problem (Duncan, 1975). The recommended solution to the unmeasured variables problem is to measure reliably all variables that are causes of an endogenous (dependent) variable and are correlated with other causes of that endogenous variable.

Regrettably, in most cases this solution is impossible to achieve, if for no other reason than that all relevant causes of an endogenous variable might not even be known (Duncan, 1975; Heise, 1975; Kenny, 1975). Consequently, the operative question is not whether one has an unmeasured variables problem but rather the degree to which the unavoidable unmeasured variables problem biases estimates of path coefficients and provides a basis for alternative explanations of results (Fisher, 1971; James & Singh, 1978). In actual practice, it is not uncommon to allow certain trade-offs, where the costs of omitting at least known causes from the causal system are evaluated in terms of their importance to the overall system and the degree to which obtained estimates of path coefficients for measured causes might be biased. Decision rules for evaluating these costs have up to now remained largely enigmatic.

This article has two objectives. The first is to summarize briefly the bases for the unmeasured variables problem and the ramifications of this problem, namely, biased solutions of path coefficients. The second objective is to provide a set of subjective decision steps that identify conditions in which an unmeasured variables problem is not likely to bias seriously the estimates of path coefficients for measured causes. In addition, several inaccuracies are noted in regard to the Billings and Wroten (1978) discussion of, and recommendations for solving, the unmeasured variables problem.

The discussion below focuses on the application of path-analytic procedures to cross-sectional, unidirectional causal models that employ nonexperimental data. Relatively

Support for this research was provided under U. S. Office of Naval Research Contract N00014-78-C-0123, Office of Naval Research Project NR 170-840, and by National Institute on Drug Abuse Grant H-81-DA-01931-01. Opinions expressed are those of the author and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy or the National Institute on Drug Abuse.

The author would like to thank Robert G. Demaree, B. Krishna Singh, and S. B. Sells for their helpful suggestions and advice, although any errors in the manuscript are the responsibility of the author.

Requests for reprints should be sent to Lawrence R. James, who is now at the School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30313

simple models are employed for illustrative purposes, and all variables are considered to be in standardized form. With the exception of the unmeasured variables problem, all other assumptions required in path analysis are assumed to be satisfied (e.g., correctly specified causal order and direction, linearity, interval scales, and no random measurement error in independent variables). In the elaboration of the argument, it was necessary to focus statistical treatments on theoretical path equations for populations. The term *solution* is employed for these treatments so as not to confuse them with *estimates* of path coefficients provided by multiple regression, that is, ordinary least squares (OLS). However, the degrees of bias represented in the solutions would be the same as those represented in population OLS estimates. (Sample OLS estimates require the addition of sampling error.)

The Unmeasured Variables Problem and Correlations With Error Terms

The existence of an unmeasured variables problem reflects a violation of an important assumption in path analysis. This assumption is that the causes for a dependent endogenous variable are uncorrelated with the error term (disturbance, residual) of the causal equation for that endogenous variable (as well as the error terms of equations for all endogenous variables that occur later in the causal order—Duncan, 1975; Johnston, 1972). This assumption implies that the causal variables in a theoretic path equation should be unrelated to unmeasured causes of the dependent endogenous variable inasmuch as the unmeasured causes are included in the error term. Satisfaction of the assumption is a necessary but not a sufficient condition for unbiased solutions of path coefficients and implies further that the error terms of different path equations in a hierarchical system of equations will be uncorrelated (Duncan, 1975).

To illustrate the issues, Figure 1a displays a path (causal) model in which X_1 is a cause of the two endogenous variables, X_2 and X_3. X_2 is also a cause of X_3. The u_1 (u_2 and u_3) are error terms that, based on the assumptions above, may have the following two components: unmeasured causal variables, which will be labeled by Zs, and random shocks (RS), which are unstable, *minor* causal influences that are generally assumed to be independent of one another. The P_{ij} are path coefficients, defined as the mean change (in standard deviation units) in a dependent endogenous variable expected to result from each unit of change in a causal variable, assuming all other causal variables in an equation are held constant (Darlington & Rom, 1972).

If it is postulated that no unmeasured

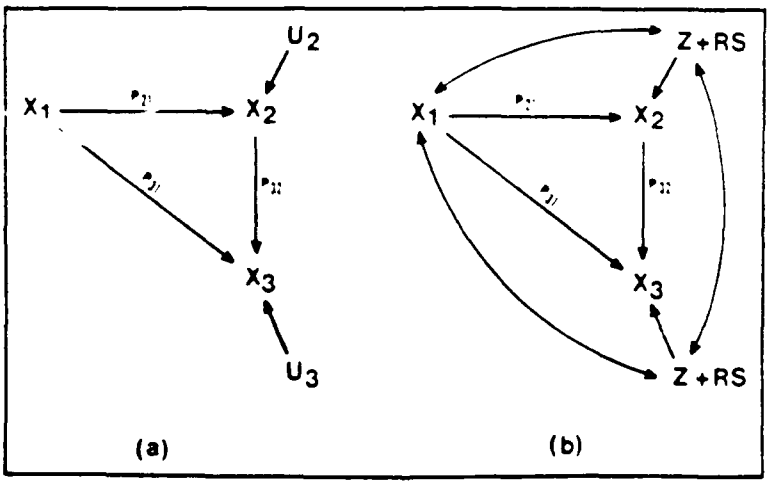


Figure 1. Illustrations of unidirectional causal models with specifications on the error terms

variables are present, then it is possible to proceed to solve for the path coefficients in Figure 1a. In this condition, the error terms would involve only the RS components, which by definition cannot be reliably measured. Thus, the assumption that the causal variables for an endogenous variable are uncorrelated with the error term for that endogenous variable would be satisfied, which connotes that X_1 is uncorrelated with u_2 and u_3, and that X_2 is uncorrelated with u_3. Note, however, that no assumption is required that X_3 be uncorrelated with u_2; that is, no assumption is required concerning relationships between errors and endogenous variables when the endogenous variables occur later in the causal order than the errors (Duncan, 1975). In this regard, Billings and Wroten (1978) are inaccurate when they stated the assumption in the following manner: "the residuals of endogenous variables are not correlated with one another or *with any other endogenous* variables" (p. 680, italics added).

Suppose that the error terms are not comprised of RS components exclusively, but rather that an unmeasured causal variable, Z, is present in both error terms. Suppose further that Z is a reliable and major cause of both X_2 and X_3 and is correlated with X_1. This state of affairs is displayed in Figure 1b, where the error terms have been decomposed into a Z component and RS components. Of initial importance is the fact that the errors will be correlated because the same Z appears in both error terms (i.e., the curved arrow between Z for X_2 and Z for X_3). This simple example demonstrates how the effects of unmeasured variables could lead directly to a violation of the assumption of uncorrelated errors.

The curved arrows from the (same) Zs to X_1 reflect correlation between Z and X_1 and connote that X_1 will be correlated with the error terms for both X_2 and X_3. Furthermore, because Z is both a cause of X_2 and is represented in the X_3 error term, it must be assumed that X_2 is correlated with the error term for X_3. Thus, all possible assumptions regarding correlations between causes and error terms may, at this time, be regarded as violated. The ramifications of this condition are discussed below.

Biased Solutions for Path Coefficients

Suppose that the path model displayed in Figure 1b is operable but that an investigator assumed incorrectly that the error terms were comprised of RS components only. The investigator could solve for the path coefficients, but they would likely be biased. To illustrate the bias, a false model (Z not included) is compared to a true model (Z included) to determine the consequences of employing the false model to solve for the path coefficients (Duncan, 1975). For example, based on Figure 1b, the path equation for X_2 in the false model is

$$X_2 = p_{21}X_1 + u_2, \tag{1}$$

in which u_2 is incorrectly assumed to be comprised of only RS components. The normal equation required to solve for p_{21} is simply

$$r_{21} = p_{21}, \tag{2}$$

which connotes that the $X_1 \rightarrow X_2$ path coefficient is equal to the zero-order correlation coefficient. The true path equation for X_2 assumes that Z is measured and is

$$X_2 = p'_{21}X_1 + p'_{2Z}Z + \mathrm{RS}_2, \tag{3}$$

in which primes are employed to designate path coefficients in the true model.

To determine the bias resulting from employment of Equation 1 rather than Equation 3 to solve for the $X_1 \rightarrow X_2$ path coefficient, we multiply through Equation 3 by X_1, take expectations, and express the results in terms of correlations. The result is

$$r_{21} = p'_{21} + p'_{2Z}r_{1Z}. \tag{4}$$

Comparison of Equation 2 with Equation 4 suggests that the use of r_{21} to solve for p'_{21} in the false model results in a bias equal to $p'_{2Z}r_{Z1}$. That is,

$$r_{21} = p_{21} = p'_{21} + p'_{2Z}r_{Z1}, \tag{5}$$

and thus p_{21} differs from p'_{21} by a factor of $p'_{2Z}r_{Z1}$.

These derivations suggest directly that p_{21} will be biased if both p'_{2Z} and r_{Z1} are greater than zero. In other words, if the unmeasured Z is a cause of X_2 and correlated with X_1, then p_{21} will be biased. If we disregard suppressors, then the bias will be in the

direction of a p_{21} that is too large; this is a direct result of failure to control for the effects of Z in solving for p_{21}.

It is extremely important to note, however, that if either p'_{2Z} or r_{Z1} is zero, or approximately zero, then little or no bias will exist in p_{21}. This suggests that bias will not occur if an unmeasured variable is in fact a cause of the dependent endogenous variable but is unrelated to the measured causes of the same variable. Consequently, it is not necessary to assume that all major causes of a dependent endogenous variable have been measured. Rather, an unmeasured cause must also be correlated with the measured causes before bias will ensue.

It is also important to recognize that there are degrees of causation: the magnitude of p_{2Z} might be anywhere on a continuum from low, to moderate, to high. Similarly, the magnitude of r_{Z1} may vary from zero, or approximately zero, to low, moderate, or high. Clearly, the product term $p'_{2Z}r_{Z1}$ may assume many permutations, only some of which are likely to result in serious bias of the solution of the path coefficient. For pragmatic purposes, it is assumed that those most likely to lead to serious bias are high–high, moderate–high, high–moderate, and moderate–moderate. Consequently, an unmeasured variables problem does not necessarily have to result in seriously biased solutions of the path coefficients. It is with the question of degree that the investigator (or critic) should be concerned. However, whenever Z is unmeasured, this is necessarily a subjective process (possible empirical procedures are addressed later).

An unmeasured variables problem will also not result in seriously biased solutions of path coefficients if an unmeasured cause is correlated highly with a measured cause. This can be demonstrated by remembering that the path coefficients involve controls for the other causal variables in a path equation. For example, consider the path equation if the unmeasured Z is included theoretically in the X_2 equation. This equation is

$$X_2 = p_{21\cdot(Z)}X_1 + (p_{2Z\cdot 1}Z) + u_2, \qquad (6)$$

in which the parentheses connote theoretical inclusions.

The path coefficient for the unmeasured Z would be approximately equal to zero if Z and X_1 were correlated highly (e.g., .95) and a control for X_1 were effected. Consequently, there is no reason to include Z in the equation because it is essentially redundant with X_1 (note also that inclusion of Z would result in a multicollinearity problem). Moreover, with Z unmeasured, essentially no bias will ensue for the p_{21} path coefficient (i.e., $p_{2Z\cdot 1}r_{Z1} \cong 0$ because $p_{2Z\cdot 1} \cong 0$).

The illustration above identifies two extremely important and related issues that should always be considered in relation to unmeasured causes. First, before an unmeasured cause is likely to create bias in the path coefficients for measured causes with which it is correlated, it must make a unique contribution to the prediction of the dependent endogenous variable. That is, it must predict meaningfully the dependent endogenous variable after controls are effected for the measured causes. Second, the preceding point can be viewed from the standpoint of *redundancy* and *linear dependence*. If a known but unmeasured cause is essentially redundant (highly correlated) with a measured cause, then there is no reason to assume that the unmeasured variable will create serious bias in the path coefficient for the measured variable. Moreover, the unmeasured cause need not simply be redundant. In more complex models involving multiple causes, it is sufficient that the unmeasured cause be essentially linearly dependent on the measured causes. A heuristic consequence of this logic is that as the number of measured causes increases, the likelihood of an unmeasured variables problem decreases. That is, even though unmeasured causes exist, they are increasingly likely to be linearly dependent, or approximately so, on the measured causes as the number of measured causes increases. Thus, it is possible to have unmeasured causes and yet have *no serious unmeasured variables problem!*

The logic developed above for a comparatively simple case of bias transfers directly to more complex cases, although in more complex cases the direction of bias may be either positive or negative. For example, serious bias in the solutions for either p_{31} or p_{32} in Figure 1b is unlikely if one of

the following conditions exists: (a) Z is only a minor cause of X_3, (b) Z is not a unique cause of X_3 (e.g., Z is linearly dependent on X_1 and X_2), or (c) Z has low correlations with X_1 and X_2. Space limitations preclude statistical development of the more complex case, and the reader is referred to an analogous development based on unstandardized variables in Duncan (1975, chap. 8).

Decision Steps for Assessing the Seriousness of Unmeasured Variables Problems

Although it is unrealistic to expect obviation of the unmeasured variables problem in research, it is possible under specified conditions to attempt to minimize bias in path coefficients to the point that the bias is within "tolerable limits" for research purposes. In the interest of identifying such tolerable limits, salient points from prior discussion are summarized below in the form of decision steps that are designed to help investigators ascertain whether an unmeasured variables problem is sufficiently serious to preclude the use of path analysis. Presentation of the decision steps must be prefaced, however, with the caution that many of the decisions require subjective judgments and the need to make empirically untestable assumptions.

The decision steps are written from the standpoint of one endogenous variable, although the steps should be applied to each endogenous variable in a causal model. Furthermore, the decision steps should be employed only when investigators have a reasonably high degree of confidence in the causal closure and stability of a causal model. However, the possibility that the model might change in the future as new causes are discovered should be clearly recognized. This is not, however, sufficient reason to preclude proceeding with causal analyses, given that no attempt is made to suggest that the present causal model is unambiguously unique or correct.

The decision steps are as follows:

Step 1. Attempt to identify known major and moderate causes of the endogenous variable.

If data have not been collected, then attempt to measure the major moderate causes, unless there appears to be a good reason not to include one or more of these variables, as determined in Step 2.

If data have already been collected, then attempt to identify known major moderate *un*measured causes. If one or more such causes is believed to exist, proceed to Step 2. If no major moderate unmeasured causes are believed to exist, then exit from the decision steps at this point (i.e., a serious unmeasured variables problem appears to be unlikely for this endogenous variable, at least from the perspective of the decision maker).

Step 2. Postulate whether each major moderate unmeasured cause is correlated with one or more of the measured causes, using prior empirical evidence whenever possible. In designing a path analysis study, this step and those that follow are meant to be viewed in terms of causes that are not as yet in the causal model, as compared to causes already included in the model.

If the correlations between an unmeasured cause and all of the measured causes are presumed to be low (e.g., 0 to $\pm$.20, although this is arbitrary), then exit here for that unmeasured cause. Note, however, that if a different unmeasured cause is included later in the causal model, then the decisions regarding prior unmeasured causes should be reevaluated; this applies to all of the following steps. Furthermore, an exit at this point suggests that the explanatory power of the causal model in regard to the endogenous variable of interest will be reduced. On the other hand, if the judgment is correct that all correlations between the unmeasured cause and the measured causes are low, then the solutions of the path coefficients for the measured causes are not likely to be seriously biased.

If an unmeasured cause is believed to have a moderate to high correlation with one or more of the measured causes, then consider whether the unmeasured cause is essentially redundant with one of the measured causes or essentially linearly dependent on some combination of the measured causes. If prior research and/or judgment allow one to have confidence in an affirmative response to one of these considerations, then exit at this point. Note again, however, that al-

though the exit suggests lack of serious bias, this will occur only if the judgments are correct.

Step 3. By reaching Step 3, it has been decided that (a) at least one unmeasured major/moderate cause exists for the endogenous variable of interest, (b) the unmeasured cause is correlated at least moderately with one or more of the measured causes, and (c) the unmeasured cause is neither redundant with one of the measured causes nor linearly dependent on some combination of the measured causes. This suggests that a serious unmeasured variables problem exists and that an attempt to solve for the path coefficients for this endogenous variable based on the measured causes is likely to result in at least one seriously biased solution. Consequently, it is recommended that path-analytic procedures not be employed for this endogenous variable until the unmeasured causes are in fact measured. (A less desirable possibility might be to delete measured causes that are presumed to be correlated with unmeasured causes.)

It should be mentioned that in a causal model involving multiple endogenous variables, it is possible to have serious unmeasured variables problems for one or more endogenous variables but not for other endogenous variables (Duncan, 1975, pp. 106–107). It is possible, therefore, to employ path-analytic procedures for only those endogenous variables without a serious unmeasured variables problem, although this is not a highly desirable state of affairs inasmuch as only part of a causal system would be addressed.

Discussion and Conclusions

In concluding, several additional points should be commented on briefly. First, when unidirectional path models are based on variables collected at only one point in time, no method is presently available to assess empirically whether an unmeasured variables problem exists, using the data at hand.[1] The controlling rule is that assumptions that moderate/major unmeasured causes are essentially uncorrelated with, or are redundant with/linearly dependent on, measured causes must be regarded as having been reasonably satisfied before OLS or other forms of estimation (e.g., maximum likelihood) are employed to estimate the path coefficients in either a population or a sample (cf. Duncan, 1975).

Second, other empirical approaches are available to assess whether an unmeasured variables problem exists and to attempt to eliminate bias created by unmeasured causes. These include time-series analysis, instrumental variables, and two-stage least squares (cf. Heise, 1970, 1975; James & Singh, 1978; Johnston, 1972; Joreskog, 1978). On the other hand, I must caution against the Billings and Wroten (1978) recommendation that rejection of the hypothesis of spuriousness in cross-lagged panel correlation (XLPC) analysis implies the absence of an unmeasured variables problem in a cross-sectional path analysis. Assume, for example, that an unmeasured Z has unique, moderate causal effects on two measured variables, X_1 and X_2. Assume further that X_1 is a moderate cause of X_2 after a control is effected for Z. In an XLPC analysis involving only the measured X_1 and X_2, the hypothesis of spuriousness (Kenny, 1975) would likely be rejected because X_1 is a moderate cause of X_2. Following the logic of Billings and Wroten, this suggests that the X_2 path equation (i.e., $X_2 = p_{21} X_1 + u_2$) does not have an unmeasured variables problem. But this is incorrect. The XLPC analysis demonstrated only that the X_1 and X_2 relationship was not completely determined by Z. In the cross-sectional path equation for X_2, if Z remains unmeasured and therefore a control for Z is not effected for p_{21}, then that path coefficient will be biased (i.e., based on the assumptions given, Z is a unique, moderate cause of X_2 and is correlated at least moderately with X_1).

Third, and finally, as discussed by Billings and Wroten (1978), and as implied in the decision steps, the unmeasured variables problem can be at least partially negated by attending first to effects and then to causes. Better yet, however, is to base the initial identification of effects and causes on a

[1] This statement should not be confused with tests of logical consistency, in which variables are in fact measured but not included in specific path equations

logical, reciprocal interaction between effects and causes. Specifically, if one wishes to examine only specific causes (e.g., leader behaviors in the context of path–goal theory), then it is desirable, if possible, carefully to refine the effects so that they reflect only the causal variables of interest. Not only will this procedure reduce the likelihood of an unmeasured variables problem, but it might also provide a much needed stimulus for more thoughtful criterion research.

References

Billings, R. S., & Wroten, S. P. Use of path analysis in industrial organizational psychology: Criticisms and suggestions. *Journal of Applied Psychology*, 1978, 63, 677–688.

Darlington, R. B., & Rom, J. F. Assessing the importance of independent variables in nonlinear causal laws. *American Educational Research Journal*, 1972, 9, 449–462.

Duncan, O. D. *Introduction to structural equation models*. New York: Academic Press, 1975.

Fisher, F. M. The choice of instrumental variables in the estimation of economy-wide econometric models. In H. M. Blalock, Jr. (Ed.), *Causal models in the social sciences*. Chicago: Aldine-Atherton, 1971.

Heise, D. R. Causal inferences from panel data. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology*. San Francisco: Jossey-Bass, 1970.

Heise, D. R. *Causal analysis*. New York: Wiley, 1975.

James, L. R., & Singh, B. K. An introduction to the logic, assumptions, and basic analytic procedures of two-stage least squares. *Psychological Bulletin*, 1978, *85*, 1104–1122.

Johnston, J. J. *Econometric methods* (2nd ed.). New York: McGraw-Hill, 1972.

Joreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrika*, 1978, *43*, 443–477.

Kenny, D. A. Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin*, 1975, *82*, 887–903.

Received June 14, 1979 ■

APPENDIX B

A STATISTICAL RATIONALE FOR RELATING SITUATIONAL VARIABLES AND INDIVIDUAL DIFFERENCES

By

Lawrence R. James
Robert G. Demaree
John J. Hater
Texas Christian University

ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 25, 354–364 (1980)

A Statistical Rationale for Relating Situational Variables and Individual Differences

LAWRENCE R. JAMES, ROBERT G. DEMAREE, AND JOHN J. HATER

Texas Christian University

Statistical rationale is presented for relating situational variables (e.g., technological complexity) to person variables (e.g., environmental perceptions, attitudes). A procedure is described wherein correlations are determined between a person variable and one or more situational variables after the scores on the situational variables have been assigned to individuals. The results of the procedure provide opportunities to ascertain (a) the degree to which variation among individuals on a person variable is associated with situational differences, and (b) the degree to which a situational variable accounts for the total possible variation in the person variable that is associated with between-group differences.

The degree to which individual differences in factors such as climate perceptions, attitudes, and behaviors are associated with differences in work situations has received increasing attention (cf., Adams, Laker, & Hulin, 1977; Herman & Hulin, 1972; Herman, Dunham, & Hulin, 1975; James & Jones, 1976; Jones & James, 1979; Lawler, Hall, & Oldham, 1974; Mowday, Porter, & Dubin, 1974; Newman, 1975; O'Reilly & Roberts, 1975; Payne & Mansfield, 1973; Payne & Pugh, 1976; Roberts, Hulin, & Rousseau, 1978; Rousseau, 1977, 1978a, 1978b; Stone & Porter, 1975). Estimates of person–situation associations are frequently based on "between-group" analyses, where membership in a particular situation (e.g., job type, work group, functional specialty, organization) is used as the independent variable (dummy variables in multiple regression, classification factors in ANOVA and multiple discriminant analysis), and scores on one or more individual difference variables, or person variables (PVs), are employed as the dependent variables. Using various forms of the general linear model, estimates of variance accounted for in the PV(s) by "group membership" (e.g., membership in different organizations) is re-

Support of this project was provided by the National Institute on Drug Abuse, Grants H-81-DA-01931-01 and R01-DA-01765, and under Office of Naval Research Contract N00014-76-C-008, Office of Naval Research Project RR042-08-01 N170-743. Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the National Institute on Drug Abuse or the Department of the Navy. The authors wish to thank Allan P. Jones, Michael K. Lindell, S. B. Sells, and John H. Wackwitz for their helpful suggestions and advice. Requests for reprints should be sent to Lawrence R. James, Institute of Behavioral Research, Texas Christian University, *Fort Worth, TX 76129*

0030-5073/80/030354-11$02.00/0

ported in the form of an eta-square, omega-square, intraclass correlation, squared multiple correlation, or multivariate analogs, such as a redundancy coefficient.

While this type of analysis reflects the amount of variation in one or more PVs associated with group membership, it is also the case that the independent variable—group (situation)—typically does not identify specific aspects of the situations represented that are associated with the variations in the PV (Firebaugh, 1979; James & Jones, 1976). This statement is perhaps more applicable to extremely general between-group designators (e.g., work group, without reference to type of work group) than to more specific between-group designators (e.g., job type or functional specialty). Nevertheless, a between-group designator such as job type is only an indirect indicator of specific situational variables, such as job complexity, role requirements, and reward structure.

Recently, emphasis has been placed on measuring specific situational variables and relating these variables to PVs (cf., Jones & James, 1979; Rousseau, 1978b). For example, in each of these studies, measures of specific, subunit situational variables (e.g., technology and centralization and formalization of structure for divisions/departments) were related to individuals' perceptions of job characteristics. The analytic procedure was also the same: all individuals in a particular division (department) were assigned the same scores on the situational variables, and then the situational scores were correlated with individuals' perceptions of job characteristics (the PVs) on the *individual sample*. It is important to note that (a) the desired level of analysis in both studies was the individual, and (b) it was assumed that the situational variables were homogeneous for all individuals in a particular division or department (see Roberts et al., 1978, pp. 106–107, for a discussion of homogeneity).

The information provided by relating specific situational variables to PVs, following assignment of the situational scores to individuals, should be superior to the information provided by between-groups analysis because the investigator now has an empirical basis for attempting to explain what it is about work environments that is associated with the PVs (James & Jones, 1976; Roberts et al., 1978). However, it is also the case that this procedure will likely result in a loss of predictive power in comparison to the between-groups procedure, which employs only group membership as a predictor. This is because the between-groups procedure identifies *all* reliable variation in a PV that is associated with between-group differences, while the use of specific situational variables generally involves only a subset of the variables that are associated with between-group differences in the PV. Thus, a salient question is: To what extent does the magnitude of the relationship between one or more situational variables and a PV approach the magnitude of the relationship between the PV and

between-group differences (for the sample studied)? In effect, this question may be viewed as one which asks for an assessment of the degree to which reliable between-group variance in a PV remains to be accounted for after the measured situational variables are considered.

The primary objective of this article is to present a statistical rationale for relating a PV to one or more situational variables and for determining the extent to which the obtained relationship approaches the (maximum) variation in a PV that is associated with between-group differences. Univariate procedures are presented initially, primarily to simplify the discussion, and are followed by an extension to the multivariate case. An empirical illustration is presented.

Statistical Rationale for Relating Situational Variables to Person Variables

For illustrative purposes, the following conditions were assumed:

(1) S_i is a continuously distributed situational variable (e.g., technological complexity), on which each of k groups (e.g., job types, work groups, departments, divisions, organizations) has a unique score ($i = 1,2, \ldots k$), although some groups may have the same score as other groups. When all individual members of the same group are assigned the same value of S_i for that group, the designation S_{ij} is used, where j represents the jth individual in a group comprised of n_i individuals ($j = 1,2, \ldots n_i$). It is assumed that the S_i for each group is homogeneous for all n_i individuals.

(2) Y_{ij} is the jth individual's score in the ith group on the person variable (PV). Note especially that the Y_{ij} are not constrained to be equal for all n_i individuals in the ith group.

When the S_{ij} and Y_{ij} are each expressed in grand-mean deviation form (s_{ij}, y_{ij}—deviations from the grand mean of all individuals across all groups), the correlation between S_{ij} and Y_{ij} can be expressed as follows:

$$r_{y_{ij}s_{ij}} = r_{ys} = \left(\sum_i \sum_j y_{ij}\, s_{ij}\right) / N \left[(1/N)\left(\sum_i \sum_j y_{ij}^2\right)\right]^{1/2} \left[(1/N)\left(\sum_i \sum_j s_{ij}^2\right)\right]^{1/2},$$

where $N = \sum n_i$. This equation may be expressed in somewhat different form by noting that all s_{ij} in group i are identical and thus $\sum_j s_{ij} = n_i s_i$ (and $\sum_j s_{ij}^2 = n_i s_i^2$), and that $\sum_j y_{ij} = n_i \bar{y}_i$. Hence,

$$r_{ys} = \left(\sum_i n_i \bar{y}_i s_i\right) / N \left[(1/N)\left(\sum_i \sum_j y_{ij}^2\right)\right]^{1/2} \qquad (1)$$

$$\left[(1/N)\left(\sum_i n_i s_i^2\right)\right]^{1/2},$$

$$= \sigma_{\bar{y}_i s_i} / (\sigma_{y_U} \sigma_{s_i}). \quad (2)$$

Equations (1) and (2) demonstrate that the correlation between S_U and Y_U on the total individual sample—r_{ys}—is a function of (a) the covariance between the weighted group means on the PV and each group's score on the situational variable, and (b) the standard deviations of the PV and the situational variable on the total individual sample. Of interest are the facts that r_{ys} will only achieve absolute values greater than zero when (a) variation exists among the group S_i scores, (b) there is comparatively more between-group variation in the Y_U's than within-group variation, and (c) the group mean Y_U's covary with the S_i. The differential weighting due to different n_i may become a confounding factor if the group n_i are substantially different (i.e., larger groups will have stronger effects on r_{ys}), and thus caution should be used when analyses employ groups with large differences in group sample sizes. Nevertheless, Eq. (2) makes clear the fact that r_{ys} reflects the extent to which group (mean) differences on a PV tend to covary with group scores on a situational variable, relative to individual variation on the PV and between-group variation on the situational variable. This connotes that r_{ys} may be interpreted as an *association* between a situational variable and a PV.

Our next concern is the degree to which r_{ys}^2 approaches the maximum variation in the PV that is associated with between-group differences. To achieve this goal, it is necessary to determine both the total amount of variation in the PV that is associated with between-group differences and that portion of this total variation that is associated with differences in S_i. The determinations of these variations are rather easily achieved by first deriving an equation for the correlation between the S_i and the weighted group means on Y_U. This correlation, designated $r_{\bar{y}s}$, is as follows:

$$r_{\bar{y}s} = \left(\sum_i n_i \bar{y}_i s_i\right) / N\left[(1/N)\left(\sum_i n_i \bar{y}_i^2\right)\right]^{1/2}$$

$$\left[(1/N)\left(\sum_i n_i s_i^2\right)\right]^{1/2},$$

$$= \sigma_{\bar{y}_i s_i} / (\sigma_{\bar{y}_i} \sigma_{s_i}), \quad (3)$$

where $\sigma_{\bar{y}_i s_i}$ is the covariance between the weighted group means on the y_U and the situational variable s_i; $\sigma_{\bar{y}_i}$ is the standard deviation of the weighted

group means on y_U; and $\sigma_{\bar{s}_i}$ is the standard deviation of the s_i in the total sample.

Comparison of Eq. (3) with Eq. (2) demonstrates that the numerators are the same, as are the standard deviations for the situational variable in the denominator. However, the remaining terms in the denominators, σ_{y_U} (Eq. (2)) and $\sigma_{\bar{y}_i}$ (Eq. (3)), can generally be assumed to be unequal given that the PV—y_U—would usually be expected to vary among individuals in the same group.

If Eqs. (2) and (3) are each solved for the covariance terms, and the solutions set equal to each other, we have the following

$$r_{y_U} \sigma_{y_U} \sigma_{s_i} = r_{\bar{y}_i} \sigma_{\bar{y}_i} \sigma_{s_i}$$

which, after solving for r_{y_U} and squaring all terms, is

$$r_{y_U}^2 = (\sigma_{\bar{y}_i}^2 / \sigma_{y_U}^2)\, r_{\bar{y}_i}^2. \tag{4}$$

Furthermore, $\sigma_{\bar{y}_i}^2 / \sigma_{y_U}^2$ is η_i^2, the squared correlation ratio (eta square) of y_U on group membership. Thus, Eq. (4) is

$$r_{y_i}^2 = \eta_i^2\, r_{\bar{y}_i}^2. \tag{5}$$

where $r_{y_i}^2$ is the proportion of the variance in a PV associated with situational variable S_i; η_i^2 is the *total* amount of variation in the PV that is associated with between-group differences; and $r_{\bar{y}_i}^2$ is the variance in the weighted group mean PV scores that is associated with differences in the situational variable S_i.

Viewed from another perspective, η_i^2 is the maximum possible variation in the PV that is associated with between-group differences. $r_{y_i}^2$ will be equal to η_i^2 only in the condition that $r_{\bar{y}_i}^2 = 1.0$, which can be seen in Eq. (5). Note that $r_{\bar{y}_i}^2$ will be less than 1.0, and therefore $r_{y_i}^2 < \eta_i^2$, when (a) the relationship between the $\bar{y}_i$ and s_i is nonlinear, and/or (b) between-group variation exists in the y_i that is not associated with s_i (see Eq. (3)). Assuming relationships to be *linear*, which can be checked empirically, we see that $r_{\bar{y}_i}^2$ represents the proportion of variation in η_i^2 that is included in $r_{y_i}^2$. In other words, $r_{\bar{y}_i}^2$ indicates the degree to which the obtained $r_{y_i}^2$ approaches the maximum possible variation in a PV associated with between-group differences.

This is seen simply by converting Eq. (5) to

$$r_{\bar{y}_i}^2 = r_{y_i}^2 / \eta_i^2 \tag{6}$$

It is important to note that meaningful interpretation of $r_{\bar{y}_i}^2$ requires careful attention to the values of $r_{y_i}^2$ and η_i^2 inasmuch as $r_{\bar{y}_i}^2$ may assume high values that are essentially meaningless. To illustrate, if $r_{y_i}^2 = .010$ and $\eta_i^2 = .011$, then $r_{\bar{y}_i}^2 = .909$. The value of .909 suggests that $r_{y_i}^2$ did in fact approach the maximum possible variation in Y associated with between-

group differences, as reflected by η_Y^2. However, an η_Y^2 of .011 indicates that essentially none of the variation in Y was associated with between-group differences in the first place. That is, the variation in Y almost exclusively is *within-group variance*, and all that the $r_{Y_i}^2$ of .909 indicates is that one has accounted for approximately 91% of, in effect, nothing.

On the other hand, η_Y^2 (and $r_{y_i}^2$) may assume reasonably high values, in which case the information provided by Eq. (6) is salient. A straightforward use of this information would be to ascertain whether additional situational variables should be added to a study in the interest of accounting for reliable variance that still remains between groups. That is, $1 - r_{y_i}^2$ indicates the proportion of between-group variation in the PV that is *not* accounted for by the situational variable S_i. If $1 - r_{y_i}^2$ is not equal to zero, then the indication is that additional situational variables are needed in the analysis.

The inclusion of additional situational variables in the analysis means that the univariate correlation analysis will be replaced with a multiple correlation analysis. The transfer to a multiple correlation paradigm is easily achieved: the preceding logic extends directly to multiple correlation analyses based on two or more situational variables which have the same values for all individuals in each group. That is, it can be shown that the squared multiple correlations are related by $R_i^2 = \eta_i^2 R_{\hat{Y}}^2$, where R_i represents the squared multiple correlation between one PV and two or more continuously distributed situational variables. Using the same logic as above, $R_{\hat{Y}}$, the squared multiple correlation between the weighted group means on the PV and the situational variables indicates the degree to which R_i^2 approaches the maximum variation in the PV that is associated with between-group differences, as reflected by η_i^2.

It is noteworthy that some portion of the between-group variation reflected by η_Y^2 might not be limited to strictly situational attributes. For example, a part of the between-group variation in the PV might reflect *group mean* differences in individual difference variables such as age, education, experience, socioeconomic status, and so forth. This suggests that $R_{\hat{Y}}$ might not achieve a value of 1.0 by adding only situational variables to the analysis. Consequently, it would be informative to ascertain whether group mean differences on individual difference variables account for between-group variation in the PV before effort is extended to identify additional situational variables to include in the analysis.

The most straightforward approach for addressing this issue is to compute group means on the individual difference variables that are believed to be related to group differences (i.e., explain *between-group variation* in the PV) and to treat the means *statistically* as if they were situational variables (i.e., assign the group means for a group to all individuals in the group). In effect, the analytic procedure would consist of regressing the

PV on both the measured situational variables and the group means of the individual difference variables (IDVs), following the procedures outlined for situational variables. If the R_{ξ}^{2} provided by this analysis is less than 1.0, then the indication is that additional situational variables are needed.

It is important to note that the use of group means on IDVs is recommended *only as a statistical heuristic* for ascertaining whether additional situational variables are needed in an analysis. This is because the group mean scores on the IDVs are, in general, "fictional variables" with respect to individual group members, and thus cannot be interpreted meaningfully as "group variables" in the analysis. To be sure, within the context of a particular theory, a group mean score on an IDV might be interpreted meaningfully and employed and interpreted just like any other situational (group) variable. However, the group mean on an IDV will generally lack theoretical import and thus should be employed only as a statistical heuristic to ascertain if additional situational variables might be included in a study.

An Illustration

To illustrate the use of the above rationale, one set of data were selected from an ongoing research study (Hater, Note 1). The data included (a) subordinates' perceptions of interdepartmental conflict (Y_{c}) on the part of 124 high level, technical personnel in an information systems department in a private health care foundation (e.g., systems analysts), (b) measures of work group centralization of decision making (S_{1i}, where the first subscript connotes situational variable number) and work group formalization of work roles (S_{2i}). Separate measures of S_{1i} and S_{2i} were obtained for each of the 19 work groups in which the 124 subordinates were employed (work group supervisors provided the S_{1i} and S_{2i} scores). A one-way ANOVA, using the 19 work groups as the independent variable (classification factor) and the perceptions of interdepartmental conflict as the dependent variable, resulted in an η_{Y}^{2} of .26 ($p < .05$). This connotes that 26% of the variance in perceptions of interdepartmental conflict was associated with between-group variations in the 19 work groups.

The squared correlations between the two situational variables and perceptions of interdepartmental conflict are presented in column one of Table 1 under univariate analysis (i.e., the r_{Yi}^{2} column). Following prior discussion, the correlations were computed by assigning each individual in group j ($j = 1, \ldots, 19$) the same S_{1i} and S_{2i} scores, and then correlating the Y_{ij} and S_{1i} and S_{2i} scores on the total (i.e., across group) subordinate sample. Before squaring, the correlations were significant and positive. The positive correlations suggest that individuals in high level technical jobs, which require a certain degree of flexibility, autonomy, and bound-

TABLE 1

RELATIONSHIPS BETWEEN SUBORDINATES' PERCEPTIONS OF INTERDEPARTMENTAL CONFLICT AND CENTRALIZATION OF DECISION MAKING AND FORMALIZATION OF WORK ROLES

Situational variable	Relationship	
Univariate analysis		
	r^2_{yi}	r^2_{yi}
Centralization of decision making (S_{i1})	.05*	.19
Formalization of work roles (S_{i2})	.07**	.27
Multiple correlation analysis		
	R^2_y	R^2_y
S_{i1}, S_{i2}	.10**	.38

Note. All analyses based on individual subordinate sample ($N = 124$).
* $p < .05$
** $p < .01$.

ary spanning, are likely to perceive a lack of cooperation and more conflicts among organizational departments when decision making processes are constrained by centralized and formalized structures (cf., James & Jones, 1976).

The r^2_{yi} column in Table 1 under univariate analysis indicates the *proportion* of total variation in subordinates' perceptions of interdepartmental conflict that was *both* (a) associated with between-group differences and (b) accounted for by either centralization or formalization (the relationships were linear). For example, centralization of decision making accounted for 19% of that variance in interdepartmental conflict that was associated with between-group differences. Consequently, 81% of the variance in the perceptions that was associated with between-group differences was *not* accounted for by centralization (i.e., $1 - r^2_{yi}$). It is important to note that r^2_{yi} need not be calculated directly. One only needs to calculate η^2_y, each r^2_{yi}, and then divide each r^2_{yi} by η^2_y (see Eq. (6)). In addition, "accounted for" is used only in the statistical sense, and does not imply causal attribution of variance.

The lower part of Table 1 presents the results of the multiple correlation analysis. Following assignment of scores to individuals, centralization and formalization were correlated .30 ($N = 124$ subjects, $p < .01$), which connotes that the values of the r^2_{yi}'s from the univariate analysis could not simply be added to obtain an estimate of the variance in Y_y associated with the combined situational variables. The squared multiple correlation, R^2_y, again computed on the subordinate sample, was .10 ($p < .01$). Division of R^2_y by η^2_y, which provided R^2_y, was .38 (i.e., .10/.26), suggesting that 38% of the variation in subordinates' perceptions of interdepartmental conflict that was associated with between-group differences was accounted for by a linear combination of centralization and formalization.

Since the relationships among the variables were linear, the results of the analysis above indicates clearly that additional, between-group predictors are needed in the study. That is, based on $1 - R_t^2$, 62% of the between-group variation in the perceptions remains to be accounted for. We believe this is worth knowing! It should also be noted that if the differences between η_v^2 and R_v^2 reflect nonlinearity, various forms of polynominal regression or moderator analysis would be indicated.

SUMMARY AND CONCLUSIONS

The primary objectives of this article were to present statistical rationales for relating a person variable to one or more situational variables, following assignment of scores on the situational variables to individuals, and for determining the degree to which the obtained relationship approaches the maximum variation in a person variable that is associated with between-group differences. It was shown that the correlation between a situational variable and a PV was a function of between-group variation on the PV, in relation to within-group variation, and covariation of the group means on the PV with the group scores on the situational variable. It was also shown that the squared correlation between a continuously distributed situational variable and a PV could be decomposed into (a) an eta square, which is the maximum variation in a PV associated with between-group differences, and (b) the squared correlation between the weighted group means on the PV and the situational variable (r_{yt}^2). This decomposition had the important implication that r_{yt}^2 reflects the degree to which the obtained r_{wt}^2 approaches the maximum variation in a PV associated with between-group differences, as measured by η_v^2.

Extensions to the multivariate case were presented, and an application of the procedures to empirical data was illustrated. Finally, as part of the process of ascertaining whether additional situational variables are needed in a study, it was recommended that group means on individual difference variables (IDVs) which help explain between-group variation in a PV be entered into the analysis. It was noted, however, that this procedure generally served only as a statistical heuristic to determine whether R_t was less than 1.0 after the group means on the IDVs had been entered into the analysis, in conjunction with the measured situational variables. Only in the case that a group mean on an IDV has theoretical relevance as a "group variable" should the mean be retained in the analysis for interpretative purposes.

Several additional points deserve mention. First, a note of caution needs to be offered concerning the number of situational variables (and group means on IDVs in the analysis described above) employed as predictors in relation to the number of groups. Ordinarily there should be many more groups than situational variables. When this is not the case, the interpretation of results must be guarded. For example, if there are

only two groups, a single situational variable whose value differs for the two groups will serve as an identifier of membership in the groups and will account fully for the between-group variation of a PV, irrespective of whatever conceptual meaning may be deserved for the situational variable. In general, if there are k-1 situational variables (where k is the number of groups), and none of these variables can be perfectly predicted linearly by one or more of the remaining situational variables, R_y^2 will always be equal to 1.00. In such a case, the set of situational variables merely serves to identify the membership in groups and will always yield $R_y^2 = \eta_y^2$ and thus $R_y^2 = 1.0$. The same would be true for a set of randomly generated situational variables (cf. Cohen & Cohen, 1975), and thus it should be clear that as the number of situational variables approaches or reaches the number of groups minus one $(k-1)$, the closeness of R_y^2 to η_y^2 has lesser relevance to the substantive import of the situational variables and more relevance to their role as identifiers of group membership. The foregoing is of little concern when the number of groups is very large in comparison to the number of situational variables, but in some studies this may not be the case.[1]

Second, with purely correlational data, it is generally not meaningful to attempt to infer that the variance attributions (η_k^2, r_{yt}^2, r_{yt}^2, R_y^2, R_t^2) are causal. For example, James, Hater, Gent, and Bruni (1978) and Roberts et al. (1978) discuss errors that evolve from making causal attributions of variance in a PV to situational variables, based on correlational data, when the true underlying causal model involves reciprocal causation between persons and situations.

Finally, with the exception of η_y^2, we have focused exclusively on continuously distributed situational variables, which reflects our bias toward the use of parametric procedures whenever possible. However, the rationale developed is equally applicable to categorical variables, where, for example, a situational variable is operationalized in terms of different types of training received. In this case, R_y^2 is determined by the use of well-known dummy variable procedures (Cohen & Cohen, 1975), or perhaps a mix of dummy variables and continuously distributed variables, and the relationship $R_y^2 = \eta_y^2 R_t^2$ is applicable.

REFERENCES

Adams, E. F., Laker, D. R., & Hulin, C. L. An investigation of the influence of job level and functional specialty on job attitudes and perceptions. *Journal of Applied Psychology*, 1977, 62, 335–343.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Wiley, 1975.

Firebaugh, G. Assessing group effects: A comparison of two methods. *Sociological Methods and Research*, 1979, 7, 384–395.

Herman, J. B., Dunham, R. B., & Hulin, C. L. Organizational structure, demographic cha-

[1] This issue is discussed in greater detail in Demaree, James, and Hater (Note 2).

acteristics, and employee responses. *Organizational Behavior and Human Performance*, 1975, 13, 206–232.

Herman, J. B., & Hulin, C. L. Studying organizational attitudes from individual and organizational frames of reference. *Organizational Behavior and Human Performance*, 1972, 8, 84–108.

James, L. R., Hater, J. J., Gent, M. J., & Bruni, J. R. Psychological climate: Implications from cognitive social learning theory and interactional psychology. *Personnel Psychology*, 1978, 31, 783–813.

James, L. R., & Jones, A. P. Organizational structure: A review of structural dimensions and their conceptual relationships with attitudes and behavior. *Organizational Behavior and Human Performance*, 1976, 16, 74–113.

Jones, A. P., & James, L. R. Psychological climate: Dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance*, 1979, 23, 201–250.

Lawler, E. E., III, Hall, D. T., & Oldham, G. R. Organizational climate: Relationship to organizational structure, process, and performance. *Organizational Behavior and Human Performance*, 1974, 11, 139–155.

Mowday, R. T., Porter, L. W., & Dubin, R. Unit performance, situational factors, and employee attitudes in spatially separated work units. *Organizational Behavior and Human Performance*, 1974, 12, 231–248.

Newman, J. E. Understanding the organizational structure–job attitude relationship through perceptions of the work environment. *Organizational Behavior and Human Performance*, 1975, 14, 371–397.

O'Reilly, C. A., III, & Roberts, K. H. Individual differences in personality, position in the organization, and job satisfaction. *Organizational Behavior and Human Performance*, 1975, 14, 144–150.

Payne, R. L., & Mansfield, R. Relationships of perceptions of organizational climate to organizational structure, context, and hierarchical position. *Administrative Science Quarterly*, 1973, 18, 515–526.

Payne, R. L., & Pugh, D. S. Organizational structure and climate. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976.

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. *Developing an interdisciplinary science of organizations*. San Francisco: Jossey-Bass, 1978.

Rousseau, D. M. Technological differences in job characteristics, employee satisfaction and motivation: A synthesis of job design research and sociotechnical systems theory. *Organizational Behavior and Human Performance*, 1977, 19, 18–42.

Rousseau, D. M. Measures of technology as predictors of employee attitudes. *Journal of Applied Psychology*, 1978, 63, 213–218. (a)

Rousseau, D. M. Characteristics of departments, positions, and individuals: Context for attitudes and behavior. *Administrative Science Quarterly*, 1978, 23, 521–540. (b)

Stone, E. F., & Porter, L. W. Job characteristics and job attitudes: A multivariate study. *Journal of Applied Psychology*, 1975, 60, 57–64.

REFERENCE NOTES

1. Hater, J. J. *Role conflict and role ambiguity as psychological climate variables*. Unpublished doctoral dissertation, Texas Christian University, 1978.
2. Demaree, R. G., James, L. R., & Hater, J. J. *A statistical rationale for relating situational attributes and individual differences*. Institute of Behavioral Research Report 79-1. Fort Worth, Tex.: Texas Christian University, Institute of Behavioral Research, 1979.

RECEIVED June 14, 1979

APPENDIX C

BEYOND AUTOREGRESSIVE MODELS: SOME IMPLICATIONS OF THE TRAIT-STATE DISTINCTION FOR THE STRUCTURAL MODELING OF DEVELOPMENTAL CHANGE

By

Christopher Hertzog
Georgia Institute of Technology

John R. Nesselroade
The Pennsylvania State University

Beyond Autoregressive Models: Some Implications of the Trait-State Distinction for the Structural Modeling of Developmental Change

Christopher Hertzog
Georgia Institute of Technology

John R. Nesselroade
The Pennsylvania State University

HERTZOG, CHRISTOPHER, and NESSELROADE, JOHN R. *Beyond Autoregressive Models: Some Implications of the Trait-State Distinction for the Structural Modeling of Developmental Change.* CHILD DEVELOPMENT, 1987, 58, 93–109. The use of structural modeling techniques to fit change concepts, including developmental ones, to repeated-measurements data has been rather firmly but uncritically wedded to autoregressive model specifications. The uncritical application of an autoregressive specification to repeated measures does not take into account subtleties of conceptions of stability and change (e.g., the trait-state distinction) that are now recognized in the behavioral research literature. We review the basic distinction between trait and state and examine the implications of the different possibilities for modeling developmental phenomena. The arguments are illustrated with empirical examples.

One of the primary arguments favoring longitudinal data is the utility of time-structured observations for explaining causal relations among variables that cannot be experimentally manipulated (see Biddle & Marlin, 1987, in this issue; Crano & Mendoza, 1987, in this issue; Dwyer, 1983; Heise, 1975). Such is the case in studies of development, in which an analysis of pertinent phenomena must proceed by observing development-in-context, usually without the opportunity to intervene in the developmental process. This state of affairs helps to explain the enthusiasm for structural regression models of longitudinal measurement evident among many developmental scientists (e.g., Nesselroade & Baltes, 1984; Schaie & Hertzog, 1985).

A very common and popular structural regression model for longitudinal data is referred to as a first-order autoregressive model, meaning a model in which variables are represented as causes of themselves over two points in time (Dwyer, 1983; Jöreskog & Sörbom, 1977; Kessler & Greenberg, 1981). These models form the basis for techniques such as cross-lagged regression analysis (Kenny, 1979; Rogosa, 1979) and have been argued to be optimal modeling techniques for studying stability and change in developmental applications (e.g., Jöreskog, 1979; Schaie & Hertzog, 1985). Hertzog (1985) reviewed these models and their utility for developmental analysis in some detail.

We have come to conclude that the rationale for the first-order autoregressive model is implicitly based on a trait conception of the variables in the model. By traits we mean relatively stable and permanent attributes. The implication of our conclusion is that first-order autoregressive models may be a poor way of representing change for nontrait phenomena (states), that is, models of relations among fluctuant attributes dependent upon temporary constellations of influences and circumstances. Thus, the thesis of this article is that developmental scientists need to differentiate more systematically two conceptions of stability and change as they bear on the modeling of longitudinal data. Subsequently, we will identify prototypic classes of attributes of individuals that pertain to the change/stability distinction.

Work on this paper was supported in part by a Research Career Development Award from the National Institute on Aging (1K04-AG00335). Send reprint requests to the first author at the School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170.

We begin our discussion with the concepts of lability and stability and the nature of the trait-state distinction. Next, we present results from longitudinal analyses of mood state variables that reveal a covariance structure among mood factors that is incongruent with a first-order autoregressive model of change. Next, the possible explanations for this incongruity from a state-oriented perspective are explored. Finally, we discuss alternative methods for examining state and trait change models and identify some of the critical features of research needed to test the trait-state distinction and its implications for developmental science.

Lability versus Stability in Longitudinal Data

Longitudinal research maintains a certain mystique, especially for developmentalists. At both the manifest- and latent-variable levels, longitudinal designs are considered essential for testing notions of stability and change (e.g., Baltes & Nesselroade, 1979). We agree, but whether a given longitudinal design contributes to the understanding of stability and change crucially depends on the validity of its use (Rogosa, in press). In evaluating the validity of a particular longitudinal design, we must recognize that stability or change is in fact an *intraindividual* (within-person) phenomenon. In fact, the face validity of a longitudinal design rests on the notion that change is fundamentally a property of the individual unit of observation.

The primary definition of stability (lack of change) in the literature is only indirectly a function of intraindividual change. Although there are multiple and more differentiated definitions of stability (change) (Kagan, 1980; Mortimer, Finch, & Kumka, 1982), the most common ones refer to unchanging mean levels over time (mean stability) and unchanging distributions of individual differences over time (covariance stability). Note that these two definitions refer to groups of individuals as a whole. Primarily due to the research traditions of a nomothetic (and trait-oriented) scientific worldview, there has been relatively little attention paid to a third type of stability—*intraindividual stability* (change within the given sampling unit). Instead, the first two types of stability are usually studied in traditional longitudinal research.

Covariance stability, or stability of individual differences, is reflected in the covariance of a variable with itself over two points in time. In structural regression models, covariance stability is often translated into a regression of a variable on itself in longitudinal data. These "autoregression" coefficients may be termed "*stability coefficients*." Covariance stability reflects the degree to which observed units show similar change patterns. Conversely, low levels of stability reflect, in Baltes and Nesselroade's (1973) term, "interindividual differences in intraindividual change." However, the magnitude of the stability coefficient depends both upon the intraindividual changes and upon the magnitude of interindividual differences. Stability coefficients can be high if (1) there are high levels of intraindividual change that are consistent across individuals, (2) if there is salient intraindividual change only in a (relatively small) proportion of the sampled units, or (3) if meaningful amounts of intraindividual change are nevertheless small relative to the magnitude of interindividual differences. Stability coefficients are, therefore, summary statements about relative change in a population of individuals. They are determined by, but should not be equated with, intraindividual stability (i.e., no change).

Given the multiple influences on the magnitude of stability coefficients, the interpretation of longitudinally observed measures as stable or changeable is not a clear-cut matter. A further complication is that attributions of stability seem to depend a great deal on the perspective of the interpreter. For example, a stability coefficient of +.60 over a period of 5 years can be interpreted as high or low, depending upon both psychometric concerns and one's theoretical orientation and expectations.

Nevertheless, longitudinal data are inherently more interesting to the student of lability and stability than are cross-sectional data because the former provide the necessary but not sufficient information for making such judgments. Cross-sectional data do not provide direct evidence of stability or lability at the intraindividual level. Rather, with cross-sectional data, inferences concerning stability or lability must rest on the putative nature of the variables that are measured. For example, in the absence of retest information, one may be far more likely to ascribe stability over lengthy intervals to general intelligence than to affective attributes.

In designing longitudinal studies, a consideration of the putative nature of the variables is, we believe, crucial to decisions regarding subsequent analyses and ultimately to interpretive clarity. It is in this light that we now discuss the trait-state distinction as a key

organizing construct in the conceptualization of developmental studies (Nesselroade, in press).

The Trait-State Distinction

The distinction between states and traits has a relatively long history that reaches back at least as far as Cicero (Eysenck, 1983). The distinction refers to two different classes of attributes for describing people. Traits, on the one hand, are attributes of individuals that are relatively stable across occasions. For example, having two eyes, practicing monogamy, or being an extrovert are traitlike attributes. States, on the other hand, comprise attributes of individuals that are relatively changeable in nature. Examples of statelike attributes include hormonal levels, diurnal fatigue, and situational anxiety. A dichotomy of trait and state may oversimplify the range of possibilities (Cattell, 1966; Nesselroade & Bartsch, 1977), but it suffices for our purposes here.

Interindividual differences.—Research flowing from the distinction between trait and state, especially various attempts to render the concepts operational, falls largely within the individual-differences tradition. Thus, the working definitions of the concepts have tended to focus on variation within and among individuals. We will try to draw the distinction more sharply in individual-differences terms.

Traits, because of their putative stability, are potentially useful for the purpose of discriminating between one individual and another without having to consider intraindividual change (e.g., "Jones is brighter than Smith"). As attributes that represent stable differences among individuals, traits that are valid predictors of other attributes, such as how one will react in a particular situation or performance at some task, provide a basis for effective, long-range prediction. Moreover, they are appropriate for inclusion in explanatory systems that involve distal as well as proximal causes.

States most commonly represent dimensions of intraindividual change and serve to discriminate one time or situation in the life of a person from another (e.g., "Wilson was so happy yesterday but today he seems to be depressed"). However, states also can represent differences among individuals at one point in time, provided the individuals' state changes are not in perfect synchrony (e.g., "Today, she was 'up' and he was 'down' ").

Generally it is certainly the case that most psychological attributes will neither be, strictly speaking, traits or states. That is, attributes can have both trait and state components (Nesselroade, in press). In the case of hormonal cascades, a given person may be of a certain type (e.g., diabetic) or may have a characteristic "set-point" in a homeostatic system. These aspects of the attribute would qualify as traits. Yet the pattern of flux would be considered the statelike part of the attribute. One might even wish to argue that intrinsic patterns of state variability are themselves traits. In this sense, extraversion might be a trait, but variations in gregariousness might be considered the statelike aspect of the trait. Work in the domain of anxiety has provided ample evidence of the utility of identifying trait *and* state components of anxious behavior (e.g., Cattell & Scheier, 1961; Nesselroade, in press; Spielberger, Gorsuch, & Lushene, 1969).

Misconceptions concerning trait and state.—At this point, we must identify and briefly disclaim common misconceptions that may be evoked by the terms trait and state (see also Nesselroade, in press). First, our usage of the term "trait" should not be construed as connoting immutable, genetically determined behavioral dispositions. The conception of trait as employed here, includes stable behavioral dispositions, such as cigarette smoking or chronic stress reaction, that *can* be modified but that typically remain stable over long periods of time. One of the defining characteristics of a trait is inertia. That is, a trait will remain the same unless and until organismic or environmental influences act to change it. Stable, unchanging environments promote stable behavioral dispositions, even if those dispositions are potentially modifiable by environmental intervention.

Second, a common misconception regarding states is that they are somehow ephemeral, unpredictable, unreliably measured, and hence, uninteresting. Negative attitudes toward studies of state phenomena probably derive in part from conceptual confusion of *stability* and *reliability* (see below), accompanied perhaps by the assumption that fluctuant attributes have little predictive validity or explanatory power.

Because of the relative lability of state differences among persons, states are not more difficult to use than traits for prediction, particularly in traditional schemes that base predictions solely on distributions of interindividual differences. To predict from a

levels requires the individual's trait score and knowledge of the form of the relationship between trait (predictor variable) and criterion. Predicting from state levels, however, requires some understanding of environmental contingencies and future environmental conditions. In essence, what one must do is capitalize on traitlike aspects of state dimensions to make predictions based on state information. For instance, to know that someone tends to get anxious in a certain situation and to know how that person responds when anxious can yield a prediction concerning what the person will do when placed in that anxiety-eliciting situation. The explanatory power of many behavioral theories might be greatly enhanced if they explicitly considered situational characteristics as they interact with individuals' psychological states. For example, Endler, Hunt, and Rosenstein (1962) incorporated these ideas in the assessment of anxiety. In the cognitive domain, research on state-dependent learning and memory phenomena indicates that mood or state at the time of encoding information can influence the nature of recall at some later time (e.g., Bower, 1981). Knowledge of a person's state at the time of learning can, therefore, enhance the predictability of his or her recall performance.

Discriminating trait and state.—The trait-state distinction underscores the idea that the differences existing among individuals at one point in time may well be a function of both stable and labile attributes. Therefore, in using covariation techniques such as structural equation modeling that capitalize on individual differences in data, one must be alert to the fact that the variation that is being analyzed potentially reflects latent variables of differing temporal characteristics such as states and traits.

Longitudinal Characteristics of State Measures

At this point, two empirical examples of longitudinally assessed characteristics of state measures will be briefly presented to resolve three common misconceptions about states that arise because of their intraindividual lability: (1) their measurement structure will be unstable, (2) they will show low internal consistency, and (3) they will not correlate with each other in a consistent manner. We believe the rectification of these misconceptions is highly germane to the utilization of longitudinal data on psychological states.

Older adults' data.—The first set of data pertinent to these issues consisted of self-report measures on five state dimensions (anxiety, stress, depression, regression, and fatigue) obtained on 111 older adults at two occasions of measurement. Approximately 1 month elapsed between the two measurement occasions. These data were reported by Nesselroade, Mitteness, and Thompson (1984), who found that the anxiety and fatigue indicators formed well-defined, positively intercorrelated latent variables.

We reanalyzed the Nesselroade et al. (1984) anxiety and fatigue data. The indicators used were: (1) the Anxiety scale of the 8-State Questionnaire (8SQ), Form A (Curran & Cattell, 1976); (2) the Anxiety scale of the 8SQ, Form B; (3) the Spielberger State Anxiety Scale (Spielberger et al., 1969); and (4) three four-item packets taken from the Fatigue scale of the 8SQ, Form A. Nesselroade et al. (1984) showed that the three Anxiety scales and the three Fatigue subscales formed latent variables of Anxiety and Fatigue, respectively. They also showed that these latent variables exhibited invariant factor loadings across the two longitudinal occasions.

The 8SQ Anxiety scales, Forms A and B, were designed to be parallel forms, having equal true-score variances and equal measurement-error variances (see Lord & Novick, 1968). The psychometric assumptions of parallelism can be translated into a set of testable hypotheses regarding the covariance structure of the measures (Jöreskog, 1971, 1974; Werts, Breland, Grandy, & Rock, 1980). The first goal of the reanalysis was to show that measurement of labile states does not imply labile measurement properties. That is, individual differences in state variables may properly be quite unstable. Such instability, however, does not imply that the measures are unreliable or invalid as measures of the psychological states. Instead, one can support the reliability and validity of the state instruments by showing that they have appropriate measurement properties while being sensitive to lability in individual differences in the underlying dimensions. The parallel forms for 8SQ Anxiety allowed us to test the following hypotheses: (1) the 8SQ measures have equal factor loadings and equal true-score variances within each longitudinal occasion, (2) the 8SQ forms have equal error variances within each longitudinal occasion, (3) the factor loadings and error variances for the alternate forms are equal across longitudinal occasions, and (4) the Spielberger State Anxiety Scale is congeneric, but not tau-equivalent, with the 8SQ Anxiety forms (see Lord & Novick, 1968, for a discussion of these different assumptions). The first two hypotheses relate to the

parallelism of the Anxiety measures at each longitudinal occasion, whereas the third hypothesis stipulates that the measurement properties are invariant across the longitudinal occasions. The fourth hypothesis implies that the Spielberger Scale, with the 8SQ Anxiety scale, will form a latent variable that accounts for all its reliable variance. Of these measurement-property hypotheses, Nesselroade et al. (1984) tested only for invariant factor loadings across longitudinal occasions.

These hypotheses can be understood by reference to Figure 1, which shows the basic model for the two latent variables originally tested by Nesselroade et al. (1984). In a preliminary analysis, we discovered that the longitudinal model could be fit best by allowing autocovariance between the residuals for the Spielberger tests and Fatigue subscales B and C. That is, we modeled a residual covariance for the Spielberger test between Time 1 and Time 2, a residual covariance for Fatigue B between Time 1 and Time 2, and a residual covariance for Fatigue C between Time 1 and Time 2. These residual covariances are depicted in Figure 1. The presence of the residual covariance for the Spielberger Scale forces us to abandon Hypothesis 4: the Spielberger Anxiety Scale is not a congeneric measure of the latent Anxiety variable in the presence of the 8SQ Forms A and B. There is a reliable but specific component of variance present in the Spielberger test that covaries with itself over time. However, the other hypotheses were strongly supported by the data (see Table 1).

Our reanalysis confirmed that Forms A and B have equal factor loadings and equal error variances and are therefore parallel forms (Jöreskog, 1971), and that these measurement properties (including the variances of the factors) did not change upon the second administration. We also found the reliabilities of the alternate forms for Anxiety to be unchanging over time. The estimated reliability coefficients of Forms A and B were .85 in this older population. Table 2 gives the parameter estimates from the final model. The upper half of the factor-covariance matrix presents the correlations among the latent factors. The latent factors have moderate and stationary correlations. The estimated autocorrelations for Anxiety and Fatigue were .63, and .72, respectively.

These autocorrelations, which reflect the stability of individual differences, reach a maximum of 1.0 when individual differences are perfectly preserved over time (Baltes, Reese, & Nesselroade, 1977; Blalock, 1970; Wheaton, Muthen, Alwin, & Summers, 1977). The autocorrelations are substantial for the

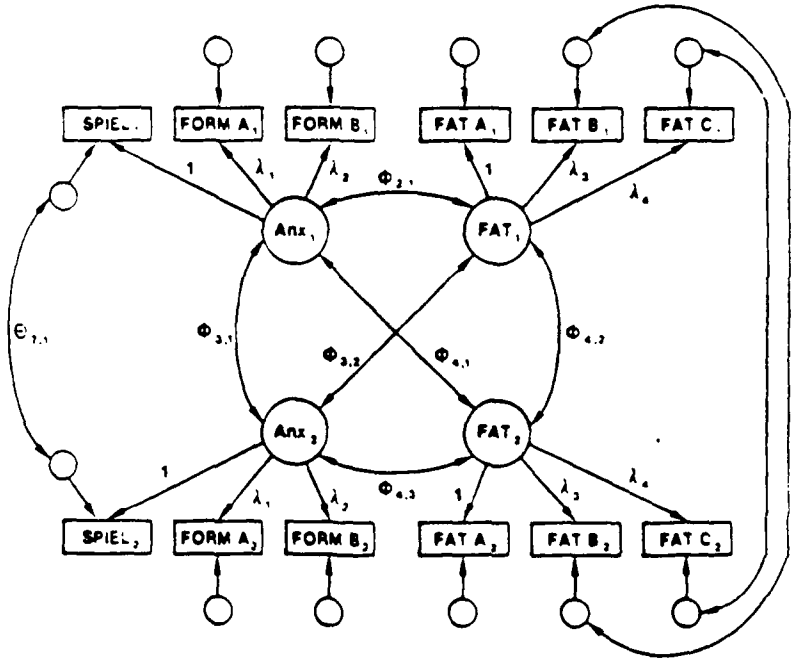


FIG. 1.—Representation of Anxiety and Fatigue interrelationships and stability as modeled by Nesselroade et al. (1984)

TABLE 1

GOODNESS-OF-FIT INDICES FOR OLDER ADULTS' MOOD STATE MODELS

Model	χ^2	*df*	*p*	GFI[a]	AGFI[b]	$\Delta\chi^2$[c]	Δdf[d]	*p*
O1. Basic model (see Fig. 1)	57.19	49	.20	.923	.88	...	...	...
O2. Tau-equivalence for Forms A and B ($\lambda_1 = \lambda_2$)	57.49	50	.22	.923	.88	.30	1	N.S.
O3. Parallelism for 8-state (over time and within time)	58.18	53	.29	.923	.89	.69	3	N.S.
O4. Parallelism and stationary latent variances ($\phi_{11} = \phi_{33}$, $\phi_{22} = \phi_{44}$)	58.50	55	.35	.922	.89	.32	2	N.S.
O5. Add stationary covariance (within occasion) ($\phi_{21} = \phi_{43}$)	58.73	56	.38	.922	.89	.23	1	N.S.

[a] LISREL goodness-of-fit index.
[b] LISREL adjusted goodness-of-fit index.
[c] Change in χ^2 from preceding model.
[d] Change in *df* from preceding model.

TABLE 2

LISREL ESTIMATES FOR FINAL MODEL ON OLDER ADULTS' MOOD STATES

	Factor Pattern Weights and Uniquenesses				
	Anxiety 1	Fatigue 1	Anxiety 2	Fatigue 2	Θ
SPIEL1	1.0*	0	0	0	55.51 (7.85)
ANXA1	2.36* (.22)	0	0	0	25.23 (2.39)
ANXB1	2.36* (.22)	0	0	0	25.23 (2.39)
FATA1	0	1.0*	0	0	1.22 (.25)
FATB1	0	.81 (.07)	0	0	4.30 (.63)
FATC1	0	.84 (.05)	0	0	1.32 (.25)
SPIEL2	0	0	1.0*	0	46.85 (6.55)
ANXA2	0	0	2.36* (.22)	0	25.23[b] (2.39)
ANXB2	0	0	2.36* (.22)	0	25.23[b] (2.39)
FATA2	0	0	0	1.0*	.55 (.25)
FATB2	0	0	0	.81 (.07)	2.61 (.40)
FATC2	0	0	0	.84 (.05)	2.40 (.37)

	Factor Covariance Matrix[c]			
	Anxiety 1	Fatigue 1	Anxiety 2	Fatigue 2
Anxiety 1	33.46[d] (7.20)[e]	.69	.63	.45
Fatigue 1	9.32[e] (1.63)	5.51[f] (.73)	.44	.72
Anxiety 2	21.03 (5.53)	5.91 (1.62)	33.46[d] (7.20)	.69
Fatigue 2	6.51 (1.63)	3.95 (.70)	9.32[e] (1.63)	5.51[f] (.73)

NOTE — * denotes fixed parameter. Abbreviations: SPIEL1, SPIEL2 = Spielberger Anxiety Scale Items 1 and 2; ANXA1, ANXB1, ANXA2, ANXB2 = 8-State Anxiety Scales, Forms A and B, at Times 1 and 2; FATA1, FATB1, FATC1, FATA2, FATB2, FATB3 = Fatigue item packets A, B, and C at Times 1 and 2.

[a] Constrained equal regression of 8-state variables on Anxiety.
[b] Constrained equal 8-state measurement error variances.
[c] Values above diagonal are factor correlations.
[d] Constrained equal Anxiety factor variances.
[e] Constrained equal covariances of Anxiety and Fatigue at Times 1 and 2.
[f] Constrained equal Fatigue factor variances.

state measures and may indicate less lability in mood states in older populations. Nevertheless, the autocorrelations do not approach the maximum of 1.0 even though a period of only 1 month separates the measurement occasions. This level of stability can be contrasted with data on psychometric intelligence in older populations reported by Hertzog and Schaie (1986), in which the autocorrelation of a general intelligence factor exceeded .9 over 7-year intervals!

The current analysis shows that the moderate levels of stability in individual differences is not a function of lack of reliability in the state measures (see also Nesselroade et al., 1984, Nesselroade, Pruchno, & Jacobs, 1985). Instead, it is attributable to lability of individual differences in latent states.

Younger adults' data.—In a second set of data, both Forms A and B of the Curran and Cattell 8SQ Anxiety Scale were administered to 42 college students at each of four occasions of measurement (Nesselroade et al., 1985). Approximately 4 days elapsed between successive measurement occasions. Although the sample size here is small for purposes of confirmatory factor analysis (see Tanaka, 1987, in this issue), the data set is didactically useful.

Our reanalysis focused again on the measurement properties of Forms A and B. Figure 2 shows the basic model originally estimated by Nesselroade et al. (1985). We used the same model on the covariance matrix of the alternate forms and tested the hypotheses of parallelism and stable measurement properties over time. Table 3 summarizes a set of models testing parallelism in Forms A and B within occasions and over time. It appears that Forms A and B have unchanging measurement properties over time, but that there is some indication that they are not perfectly parallel forms in this younger population. The reliability of the scales is high. Based on the results from Model Y3, we estimate the reliability of Form A at 1.0 and the reliability of Form B at .87. From Model Y5 (complete parallelism), the reliability of both forms is estimated at .94.

Nesselroade et al. (1985) found that the correlations among the Anxiety factors across the longitudinal occasions were quite low. We examined these correlations by testing

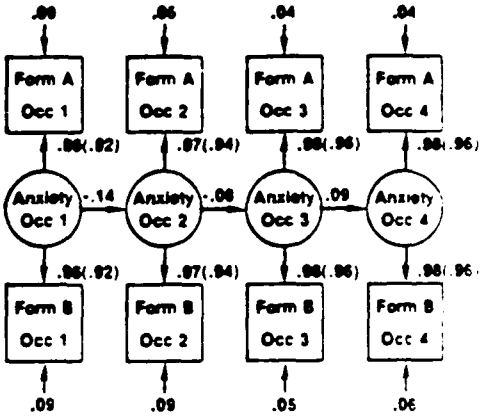


FIG. 2.—Representation of Anxiety scale reliabilities and latent variable stabilities as modeled by Nesselroade et al. (1985).

the hypothesis of orthogonal factors—that is, by requiring all factor covariances to be fixed to 0. The fit of this model, presented in the last row of Table 3, was not significantly worse than that of the preceding model of complete parallelism in error variances.

The young adults' data provide an even stronger demonstration of the differentiation of stability and reliability in state variables such as anxiety. The factor correlations of the latent Anxiety variable over time are so low that we cannot reject the hypothesis that the factors are uncorrelated in the young adult population,[1] and yet the reliabilities of the anxiety measures are high. Finally, in spite of these low covariances, there is still stationarity in the variances of Anxiety over the four occasions of measurement.

Summary.—The published literature and the reanalysis reported here present a coherent picture: factor-analytic work has demonstrated the existence of state dimensions that can be reliably measured. These state dimensions behave well when analyzed with confirmatory factor models enabling assessment of their psychometric properties. Taken together, the results of the analyses just reported suggest that the state measures have stable measurement properties over time, stationary covariance structures, and considerably less than perfect stability of individual differences. Thus, in state measures, low stability is *not* a sign of poor measurement prop-

[1] We are not suggesting the factors are orthogonal in student populations. The statistical power of this test is not high, given the relatively small sample size. The important point is that, even if the population correlations are not exactly zero, they are indeed small relative to the reliabilities of Forms A and B.

TABLE 3

GOODNESS OF FIT FOR ALTERNATIVE MODELS OF YOUNG ADULTS' ANXIETY

Model	χ^2	*df*	*p*	GFI[a]	AGFI[b]	$\Delta\chi^2$[c]	Δdf[d]	*p*
Y1. Basic model (λ for Form B equal over time)	14.11	17	.67	.93	.84	...	...	...
Y2. Tau-equivalence (all λ = 1.0)	16.06	18	.59	.92	.84	1.95	1	N.S
Y3. Equal true scores (all diagonal ϕ over time)	17.08	21	.71	.91	.85	1.02	3	N.S
Y4. Within-occasion parallelism ($\Theta_{\text{Form A}} = \Theta_{\text{Form B}}$)	27.07	25	.35	.87	.81	9.99	4	<.05
Y5. Complete Parallelism	32.50	28	.26	.85	.81	5.43	3	>.10
Y6. Parallelism *and* orthogonal factors	39.56	34	.24	.83	.82	7.06	6	N.S

[a] LISREL goodness-of-fit index
[b] LISREL adjusted goodness-of-fit index
[c] Change in χ^2 from preceding model
[d] Change in *df* from preceding model

erties of the measures but rather an indication of a high degree of lability of individuals on the underlying state dimensions. Consideration of psychological states requires scientists to select carefully research designs and techniques appropriate for assessing states and state measures. For example, test-retest coefficients are invalid reliability estimators for state measures, given lability of the states themselves. The implications of stationarity in the covariance structure of states are important with respect to the analysis of longitudinal data, as we shall now discuss.

Characteristics of Autoregressive Structural Equation Models

In this section we examine closely the underlying assumptions of autoregressive models and demonstrate that a basic first-order autoregressive model inadequately accounts for the fact that, in the Nesselroade et al. (1984) state data, the anxiety and fatigue factors maintain a moderately strong covariance with each other at the two time points.

Figure 3 shows a simple autoregressive model for two latent variables across three times of measurement (for the sake of simplicity, the measurement model is not depicted). The basic feature of the model is that each variable causes itself at the immediately following occasion of measurement. These autoregressions are the coefficients β_1, β_2, β_3, and β_4 depicted by *solid* lines. The latent variable at any occasion of measurement, call it *t*, is a function of itself at the preceding measurement occasion.

$$\eta(t) = f[\eta(t-1)]$$

This model is a first-order autoregressive model because only relationships of lag 1 ($t-1$ to t) are structured in the model. The autoregressive model depicted in Figure 3 contains two possibilities for causal influences of latent variables on other latent variables: cross-lagged regressions or simultaneous regressions. In Figure 3, the *dashed* lines represent these simultaneous (or reciprocal) influences. It should be emphasized that the model shown in Figure 3 is illustrative only; not all cross-lagged and simultaneous regressions shown can be identified and estimated. Whether one should model lagged regressions, simultaneous regressions, or some combination of the two is a matter of theory relating the timing of causal relations to the time interval in the panel design (see Kessler & Greenberg, 1981).

Let us assume for the moment that the model shown in Figure 3, including *only* the autoregressions (solid lines), is the true

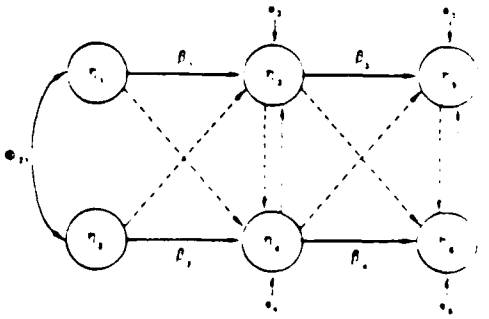


FIG. 3.—Basic structural regression model with autoregression coefficients (solid lines) and cross-regression coefficients (dashed lines).

model. The model posits an initial covariance between the two different latent variables (ϕ_{21}), but these variables do not cause change in each other over time. Hertzog (1986) referred to this model as the isolated stability model, because there is nonzero stability in each latent variable (modeled through the autoregressive coefficients), but this stability is an isolated autoregression not buttressed by cross-lagged regressions between the two latent variables. This model has been discussed by Dwyer (1983) and Rogosa (1979) as an important null hypothesis model to be rejected before alternative cross-lagged or simultaneous causal relationships between the variables can be taken seriously.

The isolated stability model is an entropic model in the sense that, in the absence of cross-lagged regressions, the covariance between the two variables will steadily decrease over time unless there is perfect stability of individual differences over time. For simplicity of exposition, we will deal for the moment with the correlations among the latent variables.[2] Assuming no omitted causes of the two latent variables, the population correlation between the two latent variables at Time 2 is

$$\rho = \phi_{21} \beta_1 \beta_2$$

At Time 3, the correlation is

$$\rho = \phi_{21} \beta_1 \beta_2 \beta_3 \beta_4$$

In general, the isolated stability model predicts decreases over time in the within-time correlations among the two latent variables unless each (standardized) β is ≥ 1; with an infinite number of occasions, the correlation decays to the entropic minimum of 0.

What then can account for the fact that in some cases within-time correlations between latent variables stay the same (as in the Nesselroade et al. data) or even increase? Mathematically, we have seen that variables in the system will become increasingly less correlated unless either (1) the correlations of the variables with themselves are 1.0 over time, or (2) through mutual causation (or causation by variables external to the system), the correlation among the variables is "built up," so to speak. The first case is one of perfect stability—nothing is changing, at least at the level of individual differences about the latent variable means. And if nothing changes, then correlations among different latent variables are preserved. But if stability of individual differences is less than perfect—that is, if there are in fact individual differences in change over time—then the correlations among the latent variables will decrease unless such changes in the two latent variables are themselves correlated due to mutual causal influence (as instantiated in cross-lagged regressions) or mutually shared causes (other than the two variables themselves).

This argument is quite complicated, so let us summarize: the implicit assumption in the autoregressive model seems to be as follows: individual differences will remain perfectly stable, and hence perfectly predictable through autoregression, unless external causes act to change the variables measured in the system. Dwyer (1983) has characterized this assumption as one of temporal inertia. The implicit corollary of this assumption is that, if stability is imperfect, there has been change in individual differences that can be modeled as a function of the causes of change. This is the apparent rationale for using the regression of the latent variable on its causes, partialed for autoregression, as a measure of the magnitude of causal influence (see Kessler & Greenberg, 1981).

Given our earlier discussion, this assumption clearly resonates with a trait conception of constructs and change in constructs. Inertia, or stability of individual differences, is expected unless other variables act to change the underlying attributes being measured. This is the basis of our concern with the standard autoregressive model for portraying longitudinally measured variables. Under the trait conception, it makes sense to assume perfect stability of individual differences unless the system is perturbed by causal influences. This assumption appears to make sense for certain psychological phenomena, that is, those suspected to be enduring such as stable attributes of individuals that have reached a determined end state (i.e., a stable individual-differences distribution). The assumption of inertial stability of individual differences modeled via autoregressive coefficients makes little sense for fluctuant attributes such as psychological states.

Recent developments in the methodological literature have demonstrated that autoregressive structural equation models should not be routinely considered *the* method of choice for analyzing change (e.g., Rogosa, in press; Rogosa & Willett, 1985). Rather, use of a

[2] The entropic nature of the model holds for covariances as well (see Dwyer, 1983).

autoregressive model must be dictated by the nature of the research question and the characteristics of the psychological phenomena under study. Rogosa and Willett (1985) have criticized the rationale of the autoregression model rather severely on the grounds that the partialed, cross-lagged regression coefficient, removing autoregression, is a poor representation of change and the causal variables' influence on change. In fact, they argue that it is often "too easy" to fit autoregressive models to longitudinal data.

In the next section we will empirically examine the usefulness of the autoregressive model with regard to the adulthood data of Nesselroade et al. (1984). Given the stationary covariance structure that we identified for the older adults' mood states, one might be optimistic that an autoregressive model with cross-lag relations will fit the data. As we shall see, this is not the case.

Fitting Autoregressive Models to the Mood State Data

Our assessment of the autoregressive model's effectiveness in modeling mood states is based solely on the older adults' data reported by Nesselroade et al. (1984). Given that we found no substantial covariance among the Anxiety factors for the undergraduate sample studied by Nesselroade et al. (1985), one could say that the autoregressive model is trivially satisfied by modeling no association in the covariance structure with autoregression coefficients of 0.

We fit a series of autoregressive models similar to that shown in Figure 3 (for two occasions of measurement only) to the older adults' data from Nesselroade et al. (1984). The measurement model for all the autoregression models used was O3 (see Table 1) specifying parallelism for the eight-state measures and correlated measurement residuals for the Spielberger and Fatigue scales. This measurement model may be considered a basis model for evaluating the fit of our autoregressive structural models. The best fit an autoregressive model could achieve is the fit of O3, which placed no constraints whatsoever on the latent covariance matrix, that is, all latent factors were allowed to covary. In Bentler and Bonett's (1980) terms, Model O3 is equivalent to a saturated model (one just identified in its structural regression equations). Therefore, we can assess the adequacy of our autoregressive models by testing their difference in fit from the basis measurement model O3 (see Hertzog, 1986).

Our first regression model specified was an isolated stability model containing autoregressions but no cross-lagged or simultaneous regression of Anxiety and Fatigue on each other. Table 4 gives the goodness of fit of this model (Model A1). It did not fit well, especially relative to the original measurement model (Model O3). This lack of fit is to be expected—and even desired—for it indicates a lack of fit to the latent covariances of a null hypothesis model of isolated stability. According to the logic of cross-lagged regression analysis, rejection of Model A1 opens the possibility that cross-lagged regressions involving Anxiety and Fatigue are required to fit the data.

The next model, A2, fitted cross-lagged regressions as well. It did not improve on the poor fit of the isolated stability model (see Table 4)! Moreover, the cross-lagged regressions were not statistically significant.

TABLE 4

GOODNESS OF FIT OF AUTOREGRESSIVE MODELS FOR OLDER ADULTS' MOOD STATES

Model	χ^2	*df*	*p*	GFI[a]	AGFI[b]	$\Delta\chi^2$[c]	Δdf[d]	*p*
Measurement model (O3 from Table 1)	58.18	53	.29	.923	.89	...	...	...
A1. Isolated stability	107.04	56	.00	.872	.82	48.86	3	<.001
A2. Cross-lagged regression	106.27	54	.00	.872	.82	48.09	1	<.001
A3. Simultaneous regressions at Time 2	68.71	54	.00	.910	.87	10.53	1	<.01
A4. Just-identified cross-lag with correlated residual, Time 2	58.18	53	.29	.923	.89	...	...	...
A5. Isolated stability with correlated residual, Time 2	58.22	55	.36	.923	.89	.04	2	N.S.

[a] LISREL goodness-of-fit index
[b] LISREL adjusted goodness-of-fit index
[c] Change in χ^2 from measurement model (O3)
[d] Change in *df* from measurement model

It is possible that the time lag is too long in these data to detect the true influences of Anxiety and Fatigue on each other using cross-lagged regressions. If so, then an obvious alternative is to specify simultaneous, reciprocal influences of Anxiety 2 and Fatigue 2 (see Dwyer, 1983; Heise, 1975). An alternative model, A3, specifying only autoregressions and the reciprocal causal influences of Anxiety and Fatigue on each other at Time 2, fared much better than Model A2 but still did not achieve the same level of fit as the measurement model. Each of these models (A2 and A3) fit poorly in spite of the fact that they have but 1 *df* in the structural equations: both estimate nine parameters (the two latent variances and the latent covariance at Time 1, four regression coefficients, and two residual variances at Time 2).

Given that we were limited in these data to two occasions of measurement, it was possible to improve the fit of the first-order autoregressive model by adding a residual covariance between Anxiety and Fatigue at Time 2. This model, A4, fit exactly as well as the original measurement model (see Table 4). This equivalent fit was no accident, however, as it was statistically determined by the fact that model A4 is just identified in the structural model. The model created 10 unique latent variances and covariances and estimated in turn 10 structural regression parameters—the latent variances and covariance at Time 1, four regression coefficients, two residual variances at Time 2, and the residual covariance at Time 2. In other words, the autoregression model was salvaged, but only by removing all restrictions on the latent covariance structure. However, both of the cross-lagged coefficients were estimated to be equal to 0! Specifically, the regression of Anxiety 2 on Fatigue 1 was estimated at .26 (SE = .29), and the regression of Fatigue 2 on Anxiety 1 was estimated to be −.003 (SE = .04). Indeed, Model A5, removing the cross-lagged regressions and specifying only the residual covariance at Time 2, provided an adequate fit to the data. Thus an autoregressive model can fit the Nesselroade et al. (1984) data, but only if we are willing to accept (1) isolated stability in autoregression and (2) the residual covariance as theoretically meaningful specifications.

How would we interpret the residual covariance? In structural regression analysis, it is common to argue for residual covariances under the assumption that there are *omitted causes* of the variables in the model that are shared between the variables (thus producing the residual covariance). In fact, Model A5 suggests that *all* relevant causes of Anxiety and Fatigue have been omitted, excepting of course the effect of each variable on itself as reflected in autoregression. At this point, however, it seems appropriate to question the need for an autoregressive model at all. This issue is considered further below.

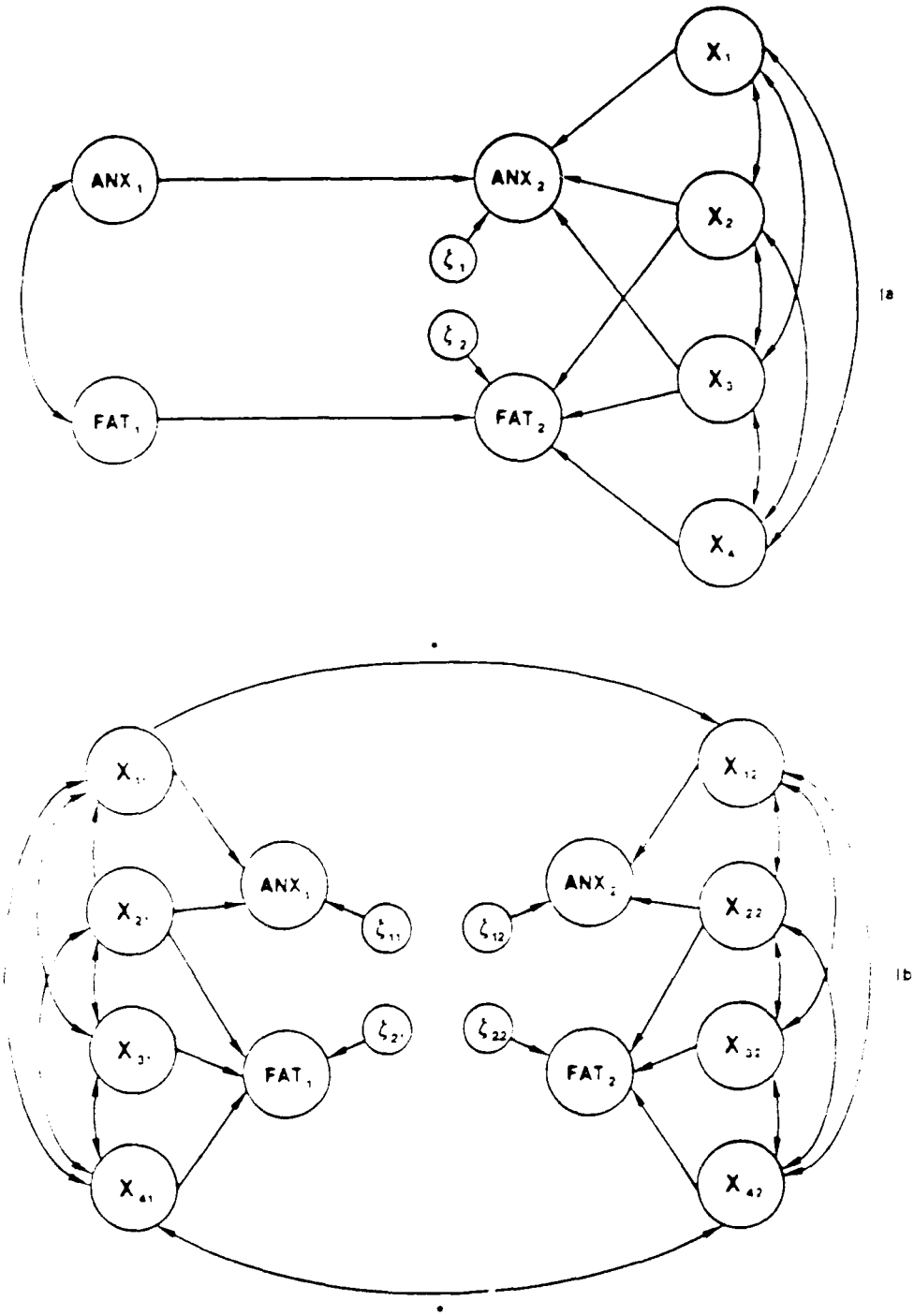
To summarize, the important conclusion from this section is that the basic cross-lagged regression model that might be thought of as the "standard" approach to modeling a two wave–two variable problem (e.g., Dwyer, 1983; Rogosa, 1979) cannot account for the stationary covariance structure we identified for the Nesselroade et al. (1984) data.

Alternative Approaches to Modeling State Phenomena

If covariance-structures approaches are to be used to model flux in psychological states, the arguments advanced here support the need to examine alternatives to conventional ways of fitting autoregressive models to panel data.

A different longitudinal panel design. — One could argue that the successful model, A5, provides an important suggestion as to the appropriate method for modeling causes of mood states. Given the salience of the correlated residual between Anxiety and Fatigue at Time 2, it appears that mutual causes of the two mood states have simply been omitted from the model. The obvious suggestion, then, is to expand the model to include the causes of the mood states at Time 2. The top panel (*a*) of Figure 4 depicts this alternative model. Given the standard rationale for the autoregressive coefficients in the model, these exogenous causal influences determine change in mood states between the two occasions. One could also expand the model to include these causes at Time 1 and model their stability over time as well.

However, our consideration of the distinction between trait and state variables calls into question the logic of assuming temporal inertia (stability) of the state variables over time. If the concerns regarding autoregressive models raised by Rogosa and colleagues (e.g., Rogosa & Willett, 1985), among others, are valid, then one should not assume that the autoregressive model is an optimal statistical method for measuring change. In that case one must consider whether it makes sense on logical grounds to argue for temporal inertia (stability) for state variables. If not, then usage of autoregressive models would appear to be



* Correlations of all X variables over time assumed but not shown.

FIG. 4.—Alternative causal models for the mood state variables. *a*, a model including autoregressions and simultaneous causes of mood. *b*, an alternative model eliminating autoregressions entirely.

contraindicated. Is it the case that one's mood at Time 1 directly causes one's mood at Time 2? The answer to this kind of question depends upon what one's theory about psychological states says about the behavior of states in the time interval between Times 1 and 2. In the case of mood states, if the interval is a matter of minutes, then perhaps there is appreciable inertia. If the time interval is a matter of months or years, then inertia per se seems unlikely. Mood states could be correlated over time, but probably not as a direct function of carryover effects from moods experienced some months prior.

This logical analysis leads us to suggest that an entirely different class of models may be needed for panel designs measuring changeable phenomena such as psychological states—namely, designs that completely eliminate autoregressive coefficients. The lower half (*b*) of Figure 4 depicts an example of this alternative modeling philosophy. The determinants of mood states are modeled as having concurrent (simultaneous) influences. The model allows for autocorrelation of mood states over the time interval, but only as a function of the correlations among the determinants of mood across time. One could, of course, posit and model autoregressive relationships among the determinants themselves, if doing so were justified on theoretical grounds.

This class of model has actually been evaluated by Hargens, Reskin, and Allison (1976) in an analysis of measurement error in panel data on scientific productivity. Analogous to the results reported here, Hargens et al. (1976) had difficulty fitting a first-order autoregressive model to yearly data on scientific productivity (as measured by variables such as the number of publications per year). Full consideration of the alternatives considered by Hargens et al. (1976) would be impossible here, but a citation of their main conclusion seems appropriate.

> Recent models for the estimation of measurement error from panel data assume a lag-1 autoregressive in the true-score variable with uncorrelated disturbances. We believe this assumption will usually be problematic for sociological variables that typically are determined by other variables having stability over time . . . we have presented a model [that] . . . assumes a first-order autoregressive process among the *disturbances* and an absence of any lagged effects in the true-score variable. This model seems particularly appropriate for variables like scientific productivity, which must be created or produced anew for each time interval, in contrast to variables that have an internal principle of stability (i.e., which tend to remain the same unless acted upon from without). [P. 457]

Clearly, the concerns raised by Hargens et al. (1976) extend beyond methods of estimating measurement error and are consistent with the arguments given here regarding the utility of autoregressive models for state variables.

We recommend that the common practice of using first-order autoregressive models for panel data be preceded by: (1) careful logical analysis of the assumptions of temporal inertia implied by this type of model, and (2) consideration of alternative models such as the one presented in Figure 4*b*. Where the endogenous variables in the panel design are seen primarily as transient states, determined by concurrent or temporally lagged, situationally-specific causes, the routine application of first-order autoregressive models may be both illogical and unwise.

Of course, psychological variables may contain both statelike and traitlike components. In such cases, a priori consideration of the existence of such state components, as well as theorizing about possible influence of transient, statelike influences on these components, may suggest panel designs in which these influences are directly measured. To scientists such as ourselves, such state components might be of central interest and a primary focus of the research. But the trait-oriented scientist would be well advised, under such circumstances, to identify and remove such components of variance from the "inertial" endogenous variables of interest. This adjustment could, in theory, be accomplished by a measurement model identifying the multiple components and their determinants (as in the multitrait-multimethod design; Joreskog, 1974) or by including the state antecedent variables in the model, as in Figure 4*b*, but retaining the autoregressive path to represent the traitlike component of the psychological variable. Failure to account for such statelike components would necessarily bias statistical estimates of causal influence and autoregressive stability.

Modeling states at the intraindividual level.—A qualitatively different alternative has been in the literature for 40 years but has recently begun to receive renewed consideration, namely, structuring the flux in states directly at the intraindividual level. Termed P technique by Cattell (Cattell, 1955, Cattell, Cattell, & Rhymer, 1947), this approach involves collecting data by assessing multiple

attributes of one individual over many occasions of measurement (see also Nesselroade, in press; Nesselroade & Ford, 1985).

The covariance matrix generated by P-technique data represents the covariation of occasion-to-occasion changes in different attributes of the individual. It can be analyzed by confirmatory factor analysis. Latent variables that are identified by such procedures manifest, by definition, coherent intraindividual variability (lability), in the sense that such lability is consistent over multiple indicators or the latent variable. Under some circumstances (e.g., Cattell, 1966), occasion-by-occasion scores on these latent variables (factors scores) can be estimated and subjected to further analysis.

P-technique, due to its direct focus on intraindividual change, provides data for modeling "steady-state" variability in the organism and both temporary and permanent changes in steady-state variability. By combining P-technique with the group design orientation in the form of concurrent P-technique studies of several individuals, one can capitalize on the strengths of both idiographic and nomothetic approaches to the study of developmental change (Nesselroade & Ford, 1985; Zevon & Tellegen, 1982). Examination of interindividual similarities and differences in the characteristics of the latent variables provides a basis for answering some important questions concerning the nature of generalizability of intraindividual change patterns over the facets of individuals and occasions (Nesselroade, 1983).

Historically, P-technique data have been modeled primarily by means of simple factor-analysis procedures. Although the results of such analyses have proven to be psychologically interesting and meaningful (Cattell & Scheier, 1961; Roberts & Nesselroade, in press; Zevon & Tellegen, 1982), the practice has been criticized (Anderson, 1961; Holtzman, 1962; Molenaar, 1985) because it does not account for the possibility of autocorrelations of variables in time series (sometimes termed "nonindependence" in the time-series literature). However, recent developments (McArdle, 1982; Molenaar, 1985) appear to provide the means for treating such statistical problems. It is our hope that this class of models for single subject behavior will enable researchers to structure the intraindividual lability inherent in states as a complementary and viable alternative to the expanded panel designs shown in Figures 4*a* and 4*b* for modeling interindividual variations in state variables. If within-person variability can be first structured at the individual (idiographic) level by multivariate analysis techniques and then examined for between-person differences and similarities, it opens a promising alternative to the study of generalizability across individuals and the construction of nomothetic relationships (Nesselroade & Ford, 1985; Zevon & Tellegen, 1982).

Summary and Conclusion

The analysis of the developmental process can be approached in a number of ways. Statistical analysis of covariance structures is one set of important techniques for this purpose, as this special issue suggests. In this article we have delineated the distinction between trait and state dimensions and its implications for the statistical modeling of longitudinally measured behavioral attributes.

First, state measures behave lawfully. They can manifest desirable measurement properties of reliability and validity while reflecting a considerable amount of lability of score at the intraindividual level. Such lability runs counter to conventional, trait-oriented conceptions of measurement and models of development. However, it cannot be dismissed as merely "error of measurement."

Second, the possibility must be recognized that the individual differences measured at any given occasion can represent labile characteristics as well as the more stable, traitlike attributes. The failure to recognize and model this possibility can lead to biased estimates of the parameters of traitlike attributes, including stability of the latent construct and reliability of its operational expressions, thus clouding the description and interpretation of data and related inferences about the nature of change.

Finally, our article has questioned the validity of standard autoregression models for change in psychological states. There is no doubt that autoregressive models will continue to fit many kinds of developmental phenomena—namely, development of psychological traits. When that happy circumstance occurs, there may be no reason to downplay their importance as descriptive representations of a temporal process (but see Rogosa & Willett, 1985).

What is at doubt is the universal validity of autoregressive models representing change over time in behavioral data. We have argued that dimensions along which individual differences are displayed are not homogeneous

and uniform. Two important and related ways that variables differ are: (1) temporal characteristics, and (2) antecedents of change. The assumptions regarding these dimensions inherent in traditional autoregressive models appear to be more applicable to variables characterized by high stability and temporal inertia (traits) rather than variables with low stability and high degree of situational and temporal specificity (states).

Explicit recognition of the differing temporal characteristics implied by the trait-state distinction serves to warn us that sole reliance on the traditional, trait-oriented concepts of differential psychology will not necessarily lead to an accurate portrayal of extant differences among individuals. Rather, an understanding of how and why individuals differ from one another requires attention both to dimensions of intraindividual variability and to the antecedents of intraindividual variability. Obviously, to account for both the transient and the more stable components of individual differences will require more complex models and procedures than envisaged in trait-oriented approaches. Nevertheless, the difficulties engendered should be more than offset by gains in our understanding of developmental processes. Our two suggestions—first, for expanded autoregressive models that include state antecedent information *and* allow for trait components to be estimated, and, second, for direct modeling of intraindividual change using single-subject designs—are offered as steps toward increasing our capacity to model stability and change.

References

Anderson, T. W. (1961). *The use of factor analysis in the statistical analysis of time series*. Technical Report No. 12, Contract AF 41, School of Aviation Medicine, Brooks Air Force Base.

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 219–252). New York: Academic Press.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York: Academic Press.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks/Cole.

Bentler, P. B., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Biddle, B. J., & Marlin, M. M. (1987). Causality, confirmation, credulity, and structural equation modeling. *Child Development, 58*, 4–17.

Blalock, H. M. (1970). Estimating measurement error using multiple indicators and several points in time. *American Sociological Review, 35*, 101–111.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*, 129–148.

Cattell, R. B. (1952). The three basic factor analytic research designs—their interrelations and derivatives. *Psychological Bulletin, 49*, 499–520.

Cattell, R. B. (1966). Patterns of change: Measurement in relation to state-dimension, trait change, lability, and process concepts. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 355–402). Chicago: Rand McNally.

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psycho-physiological source traits in a normal individual. *Psychometrika, 12*, 267–288.

Cattell, R. B., & Scheier, I. H. (1961). *The meaning and measurement of neuroticism and anxiety*. New York: Ronald.

Crano, W. D., & Mendoza, J. L. (1987). Maternal factors that influence children's positive behavior: Demonstration of a structural equation analysis of selected data from the Berkeley Growth Study. *Child Development, 58*, 38–48.

Curran, J. P., & Cattell, R. B. (1976). *Handbook for the eight-state questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. New York: Oxford University Press.

Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. (1962). An S-R inventory of anxiousness. *Psychological Monographs, 76*(Whole No. 536).

Eysenck, H. J. (1983). Cicero and the state-trait theory of anxiety: Another case of delayed recognition. *American Psychologist, 38*, 114.

Hargens, L. L., Reskin, B. F., & Allison, P. D. (1976). Problems in estimating measurement error from panel data: An example involving the measurement of scientific productivity. *Sociological Methods and Research, 4*, 439–458.

Heise, D. R. (1975). *Causal analysis*. New York: Wiley.

Hertzog, C. (1986). *On the utility of structural regression models for developmental research*. Unpublished manuscript.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intellectual development: 1. Analysis of longitudinal covariance structures. *Psychology and Aging*, 1, 159–171.

Holtzman, W. H. (1962). Methodological issues in P-technique. *Psychological Bulletin*, 59, 248–256.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, 36, 109–133.

Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: W. H. Freeman.

Joreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–352). New York: Academic Press.

Joreskog, K. G., & Sorbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D. S. Aigner & A. S. Goldberger (Eds.), *Latent variables in socioeconomic models* (pp. 285–325). Amsterdam: North-Holland.

Kagan, J. (1980). Perspectives on continuity. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 26–74). Cambridge, MA: Harvard University Press.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change*. New York: Academic Press.

Lord, F. M., & Novick, M. N. (1968). *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

McArdle, J. J. (1982). *Structural equation modeling of an individual system: Preliminary results from "A Case Study of Episodic Alcoholism."* Unpublished manuscript, Psychology Department, University of Denver.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, 50, 181–202.

Mortimer, J. T., Finch, M. D., & Kumka, D. (1982). Persistence and change in development: The multidimensional self-concept. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 4, pp. 264–313). New York: Academic Press.

Nesselroade, J. R. (1983). Temporal selection and factorial invariance in the study of development and change. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 5, pp. 59–87). New York: Academic Press.

Nesselroade, J. R. (in press). Some implications of the trait-state distinction for the study of development across the life span: The case of personality. In P. B. Baltes, D. L. Featherman, & R. M. Lerner (Eds.), *Life-span development and behavior* (Vol. 8). Hillsdale, NJ: Erlbaum.

Nesselroade, J. R., & Baltes, P. B. (1984). From traditional factor analysis to structural-causal modeling in developmental research. In V. Sarris & A. Parducci (Eds.), *Experimental psychology in the future* (pp. 267–287). Hillsdale, NJ: Erlbaum.

Nesselroade, J. R., & Bartsch, T. W. (1977). Multivariate experimental perspectives on the construct validity of the trait-state distinction. In R. B. Cattell & R. M. Dreger (Eds.), *Handbook of modern personality theory* (pp. 221–238). Washington, DC: Hemisphere/Halstead.

Nesselroade, J. R., & Cable, D. G. (1974). "Sometimes it's okay to factor difference scores"—the separation of trait and state anxiety. *Multivariate Behavioral Research*, 9, 273–282.

Nesselroade, J. R., & Ford, D. H. (1985). P-technique comes of age: Multivariate, replicated, single-subject designs for research on older adults. *Research on Aging*, 7, 46–80.

Nesselroade, J. R., Mitteness, L. S., & Thompson, L. K. (1984). Short-term changes in anxiety, fatigue, and other psychological states in older adulthood. *Research on Aging*, 6, 3–23.

Nesselroade, J. R., Pruchno, R., & Jacobs, A. (1985). Reliabilität und Stabilität in der Messung psychologischer *states*: Eine Illustration mit Massen der Angst. *Psychologische Beitrage*.

Roberts, M. L., & Nesselroade, J. R. (in press). Intraindividual variability in perceived locus of control in adults: P-technique factor analyses of short-term change. *Journal of Research in Personality*.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263–302). New York: Academic Press.

Rogosa, D. (in press). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research*. New York: Springer.

Rogosa, D., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50, 203–228.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2d ed., pp. 61–92). New York: Van Nostrand Reinhold.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R.

(1969). *The State-Trait Anxiety Inventory (STAI) test manual, Form X*. Palo Alto, CA: Consulting Psychologists Press.

Tanaka, J. S. (1987). "How big is big enough?" Sample size and goodness-of-fit in structural equation models with latent variables. *Child Development*, 58, 134–146.

Werts, C., Breland, H. M., Grandy, J., & Rock, D. R. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement*, 40, 19–29.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology, 1977* (pp. 84–136). San Francisco: Jossey-Bass.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, 43, 111–122.

Zuckerman, M. (1983). The distinction between trait and state is not arbitrary: Comment on Allan and Potkay's "On the arbitrary distinction between traits and states." *Journal of Personality and Social Psychology*, 44, 1083–1085.

APPENDIX D

STABILITY AND CHANGE IN ADULT INTELLIGENCE: I. ANALYSIS OF LONGITUDINAL CONVARIANCE STRUCTURES

By

Christopher Hertzog
K. Warner Schaie
Pennsylvania State University

Psychology and Aging
1986, Vol. 1, No. 2, 159-171

0882-7974/86/$00.75

Stability and Change in Adult Intelligence: 1. Analysis of Longitudinal Covariance Structures

Christopher Hertzog and K. Warner Schaie
Pennsylvania State University

We address two questions of central interest in adult intellectual development: the equivalence of psychometric tests' measurement properties at different ages, and the stability of individual differences in intelligence over time. We performed a series of longitudinal factor analyses using the LISREL program to model longitudinal data from Schaie's Seattle Longitudinal Study. The results indicate complete invariance in the loadings of five subtests of Thurstone's Primary Mental Abilities battery on a general intelligence factor. Individual differences in general intelligence were highly stable over 14-year epochs, with standardized factor correlations averaging about .9 between adjacent 7-year testing intervals. These results indicate that most individuals in this relatively select longitudinal sample maintained their relative ordering in intelligence.

One of the central questions in adult development regards the stability of adult intelligence—does intelligence decline with age, and if so, what is the magnitude of individual differences in patterns of change (e.g., Botwinick, 1977; Horn & Donaldson, 1980; Schaie, 1983)? The debate in the literature on the development of intelligence during adulthood has focused primarily on the stability of mean levels of intelligence—is there indeed decline, on average, on different intellectual abilities, and if so, what is the magnitude of such decline (e.g., Baltes & Schaie, 1976; Horn & Donaldson, 1976; Schaie & Hertzog, 1983)? The attention paid to stability of mean levels of intelligence has perhaps diverted the field from focusing on a different, critical—and in some senses more critical—type of stability: *stability of individual differences* in intelligence. How large are individual differences in magnitudes of age changes in intelligence during the adult years? Some developmental psychologists have suggested that adult development is characterized by increasing heterogeneity and by substantial individual differences in patterns of age change in intelligence and other cognitive capacities and skills (e.g., Baltes, Dittmann-Kohli, & Dixon, 1984; Hertzog, 1985; Schaie, 1983). Enhancement of optimal intellectual development through intervention (e.g., Schaie & Willis, 1986) requires as a first step the identification of differential patterns of aging and the isolation of the causes of such differences.

Measuring stability of individual differences in intelligence is somewhat more complex than measuring mean level stability. Although sequential sampling strategies using repeated, independent cross-sectional samples can be used to assess mean level stability (e.g., Schaie, 1977; Schaie & Hertzog, 1982), stability of individual differences can only be addressed by following individuals in a longitudinal panel design. Cross-sectional designs can only measure magnitudes of individual differences—as indicated by the variances—at a single point in time. At any given point in time, individual differences can be conceptualized as being determined by an earlier individual differences distribution and by subsequent individual differences in developmental change (see Baltes, Reese, & Nesselroade, 1977). Only a longitudinal design, by directly measuring change at the level of the individual, can be used to estimate the proportion of individual differences due to individual differences in change during preceding time periods (see Hertzog, 1985; Nesselroade & Labouvie, 1985; Schaie & Hertzog, 1985).

This study was designed to provide a careful and detailed examination of individual differences in intellectual change during adulthood. It also focuses on a second, critical issue identified by developmental methodologists regarding the assessment of change over time in variables such as intelligence. The issue is whether the constructs under study, and the measures of those constructs, are actually isomorphic at different ages. Can we assume that intelligence is the same construct at ages 25 and 75? Even if intelligence is unchanging, or continuous (Kagan, 1980) across the adult life span, is it the case that psychometric measures of intelligence are equally reliable and valid as measures of intelligence at different ages? Baltes and Nesselroade (1970) identified this issue as one of *measurement equivalence*—can we assume invariant measurement properties of empirical measures at different parts of the life span (see also Eckensberger, 1973)? As Baltes and Nesselroade indicated (see also Schaie, 1977; Schaie & Hertzog, 1985), the optimal method for assessing measurement equivalence is comparative factor analysis, in which the invari-

This article reports data collected as part of the Seattle Longitudinal Study, which has been supported over an extended period of time by grants from the National Institutes of Health, the National Institute for Child and Human Development, and the National Institute on Aging. Our work is currently supported by Grant R01-AG4770 from the National Institute on Aging.

Our thanks to William Meredith for advice and comments on our statistical models and results, and to an anonymous reviewer for helpful editorial suggestions. The cooperation and support from members and staff of the Group Health Cooperative of Puget Sound is gratefully acknowledged.

Correspondence concerning this article should be addressed to Christopher Hertzog, who is now at the School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332, or to K. Warner Schaie at S-110 Human Development Building, The Pennsylvania State University, University Park, Pennsylvania 16802.

ance of the factor structure of the psychometric abilities is assessed. As discussed elsewhere (e.g., Cunningham, 1978; Schaie & Hertzog, 1982, 1985), the best approach to the invariance problem involves the use of confirmatory factor analytic methods to test the hypothesis of age-related invariance in the factor structure.

This is the first in a series of articles describing our use of covariance structures methods to analyze patterns of change and stability in adult intelligence using data from Schaie's Seattle Longitudinal Study (SLS). In this article we describe results from a longitudinal factor model that may be used to assess (a) the measurement equivalence of the Thurstone Primary Mental Abilities battery used in the SLS and (b) the extent to which individuals in the SLS vary in patterns of intellectual change during the adult years. The Primary Mental Abilities test was developed by Thurstone and Thurstone (1941, 1949) to measure factorially pure, but intercorrelated, intellectual abilities. Assessment of factorial invariance and stability of individuals with the Primary Mental Abilities is particularly relevant, given the influence of Thurstone's work on the field of psychometric intelligence. Our findings strongly support the measurement equivalence of the Thurstone battery across much of the adult life span. We also show that there is a surprising degree of stability of individual differences in intelligence in participants from the kind of long-term longitudinal sample obtained in the SLS.

Our conclusions are based on results from a set of relatively complex longitudinal covariance structures models of the type developed by Joreskog and co-workers (e.g., Jöreskog & Sörbom, 1977). The longitudinal factor model developed by Jöreskog and others (Jöreskog, 1979; Joreskog & Sörbom, 1977) may be viewed as a generalization of other longitudinal factor analysis (e.g., models by Corballis, 1973; Corballis & Traub, 1970). To set the stage for our report, we must first summarize the methodological features of these models and how their parameters may be used to assess stability and change in individual differences over time (see also Hertzog, in press; Horn & McArdle, 1980; Schaie & Hertzog, 1985).

Let us assume that an investigator has collected multiple measures of one or more latent variables in a longitudinal design. The measures may or may not be identical at each longitudinal measurement occasion, although in the SLS the same measures were collected at each time of measurement. The relations among these variables must be represented by the covariance matrix of the observed variables (a correlation matrix should *not* be analyzed; Jöreskog & Sörbom, 1977). Given this kind of replicated longitudinal design, confirmatory factor analysis may be used to specify and estimate a longitudinal factor model with the following features.

First, the same factor structure is hypothesized to exist at each longitudinal measurement occasion. This structure is represented in the *factor pattern* matrix, which contains the regression coefficients mapping variables on factors (factor loadings). In the analysis we report here, a general intelligence (*g*) factor was modeled at each longitudinal occasion. The factors thus specified in a longitudinal factor model are often termed *occasion-specific* factors.[1] In addition to the factor pattern matrix, the basic longitudinal model includes a *factor covariance* matrix, describing the relations among the factors within and between longitudinal occasions, and a *residual covariance* matrix. The primary parameters of interest are the factor loadings and the factor covariance matrix. The first step involves evaluation of the measurement equivalence of the observed variables. Measurement equivalence may be assessed by (a) evaluating the adequacy of the model postulating isomorphic occasion-specific factors (i.e., the same number of factors with the same configuration of factor loadings at each longitudinal occasion) and (b) determining the plausibility of a model constraining these factor loadings to be equal (invariant) over all longitudinal occasions. These factor loadings are raw-score (unstandardized) regression coefficients, and invariance of these coefficients (sometimes termed *metric invariance;* see Horn, McArdle, & Mason, 1984) implies unchanging relations of the observed variables to the factors (Meredith, 1964; Schaie & Hertzog, 1985). Procedures for assessing the fit of these models are described later in the article.

Given that the hypothesis of measurement equivalence is tenable, the second step in the longitudinal analysis shifts attention to the factor covariance matrix. The diagonal elements of this matrix—the factor variances—reflect the magnitude of individual differences at each longitudinal occasion. Changes in factor variances would therefore reflect changes in the overall magnitude of individual differences over time. The stability of individual differences across longitudinal occasions is reflected in the covariances of factors with themselves over time. If the covariance of a factor at Time 1 with itself at Time 2 is large and positive, then individuals are preserving their relative order about the factor mean between Times 1 and 2. On the other hand, a zero or near zero covariance would reflect a high degree of flux in individual differences between Times 1 and 2. As shown by Baltes, Reese, and Nesselroade (1977), a zero covariance would be consistent with large individual differences in the patterns of developmental change during that time period.

Given that the SLS is a sequential study, in which multiple longitudinal samples have been followed over time (see Schaie, 1979, 1983), it is possible to expand the longitudinal model to consider longitudinal changes in multiple age groups. The extension of the model to multiple group analysis has been described by Joreskog and Sörbom (1980), and is relatively straightforward. The advantage of a multiple groups analysis in the present context is that it allows us to address the issue of age invariance in factor structure both longitudinally, within a group of individuals, and comparatively, across multiple age groups. The longitudinal samples we analyze include adults of a wide span of chronological ages who have been tested three times over a 14-year period. These multiple samples allow us to examine longitudinal invariance in factor structure over 14-year epochs, while also examining factorial invariance over the adult life-span by comparing the factor structures of multiple age groups.

Method

Subjects

The subjects in this study were participants in the Seattle Longitudinal Study conducted by Schaie and associates (Schaie, 1979, 1983). The population consisted of members of a health maintenance organization (HMO) in the greater Seattle, Washington, area. To minimize the prob-

[1] The model can be extended without difficulty to include different numbers of common factors at each longitudinal occasion, but that approach is unnecessary in our analysis.

ability of selection differences over time, the population was defined as all members of the organization as of 1956, the initial year of the longitudinal study. All participants were unpaid volunteers who answered questionnaires and took part in a psychometric testing conducted in a single session. The volunteers were recruited from a randomly drawn sampling frame of the HMO membership, stratified by age and gender. The participants were adults spanning the age range from 20 through 74, at first test, and representing a range of socioeconomic and ethnic groups. However, probability sampling was *not* employed, and the sample was therefore not necessarily representative of the entire HMO population. As was generally true of the Seattle population circa 1956, the sample is predominantly Caucasian and, reflecting the membership of the HMO, contains a higher proportion of middle- and upper-income individuals than did the total Seattle population. Further details on the population and sampling procedures may be found in Schaie (1979, 1983).

Sequential Sampling Design

The longitudinal samples studied here are a subset of the sequential samples collected in the SLS. Briefly, the design of the SLS consisted of repeated sampling from the population at 7-year intervals, beginning in 1956 and continuing through 1984. Each year of testing, a new cross-sectional sample was drawn from the population, and all previously tested individuals were contacted and recruited for participation in the longitudinal panel. Thus, each independent cross-sectional sample was transformed into a multiple-cohort longitudinal sequence (Baltes et al., 1977) by repeated testing of the same individuals. We restrict our analysis here to two 14-year longitudinal samples. Sample 1 consists of 162 subjects tested in 1956, 1963, and 1970, and Sample 2, 250 subjects tested in 1963, 1970, and 1977. The data from the two longitudinal sequences were partitioned into a hybrid sequential data matrix given in Table 1. This partition created three age groups (young, middle aged, and old) for simultaneous analysis. These age groups were formed under the assumption of no cohort differences in factor structure. Although it would have been desirable to test for both age-related and cohort-related measurement equivalence, sample sizes were insufficient for such purposes. Age-related changes in factor structure seemed more likely, a priori, and earlier work supported the assumption of no cohort differences in factor structure (Cunningham & Birren, 1980). As can be seen from Table 1, data from different birth cohorts were pooled to obtain the age groups.

Table 1
Reparameterized Sequential Sample for Multiple Group Analysis

Sample	Cohort (mean birth year)	Mean age 0_1	Mean age 0_2	Mean age 0_3	n
Group 1		30.	37.	44	109
1	1931	25.	32.	39	21
1	1924	32.	39.	46	26
2	1938	25.	32.	39	22
2	1931	32.	39.	46	40
Group 2		42.	49.	56	160
1	1917	39.	46.	53	27
1	1910	46.	53.	60	32
2	1924	39.	46.	53	51
2	1917	46.	53.	60	50
Group 3		58.	65.	72	143
1	1903	53.	60.	67	28
1	1896	60.	67.	74	15
1	1889	67.	74.	81	13
2	1910	53.	60.	67	48
2	1903	60.	67.	74	18
2	1896	67.	74.	81	21

Note 0_1 = first occasion of measurement; 0_2 = second occasion of measurement; 0_3 = third occasion of measurement.

Variables

As part of a larger psychometric battery, all of the subjects were administered the 1948 version of the SRA (Science Research Associates) Primary Mental Abilities (PMA) test, Form AM 11–17 (Thurstone & Thurstone, 1949). The 1948 PMA includes five subtests, all of which are timed and have significant speed components in adult samples (Schaie, Rosenthal, & Perlman, 1953). They are (a) Verbal Meaning—a test of recognition vocabulary, (b) Space—a test of spatial orientation requiring mental rotation in a two-dimensional plane, (c) Reasoning—a test of inductive reasoning requiring recognition and extrapolation of patterns of letter sequences, (d) Number—a test of the ability to solve simple two-column addition problems quickly and accurately, and (e) Word Fluency—a test of the ability to retrieve words from semantic memory according to an arbitrary syntactic rule. Scoring protocols followed the PMA manual. Verbal Meaning and Reasoning are scored in terms of the number of correct responses. Space and Number are scored by subtracting commission errors from the total number correct; and Word Fluency is scored by tallying the total of unique, admissible words generated.

Statistical Procedures

All of the models described were tested using the LISREL V program of Joreskog and Sorbom (1981). The analyses reported in this article used only one of LISREL's two-factor analysis measurement models. In LISREL notation, the measurement model may be specified as

$$x = \Lambda\xi + \delta, \tag{1}$$

which in matrix form specifies a q-order vector of observed variables, x, as a function of their regression on n latent variables (factors) in ξ, with regression residuals δ. The $q \times n$ matrix Λ contains the regression coefficients (factor loadings). Equation 1 implies that the covariance matrix of the observed variables in the populations, Σ, may be expressed as

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \tag{2}$$

where Λ is as before, Φ is the covariance matrix of the ξ, and Θ is the covariance matrix of the δ. Equation 2 is a restricted factor analysis model that can be extended to multiple groups (Joreskog, 1971).

The parameters of LISREL's restricted factor analysis model are estimated by the method of maximum likelihood, provided that a unique solution to the parameters has been defined by placing a sufficient number of restrictions on the equations in Equation 2 to identify the remaining unknowns. Restrictions are specified by either (a) fixing parameters to a known value a priori (e.g., requiring that a variable is unrelated to a factor by fixing its regression in Λ to 0) or (b) constraining a set of two or more parameters to be equal. The equality constraints may be applied to any subset of parameters within or between groups, which provides the basis for specifying a model requiring invariant factor structures between multiple groups or across longitudinal occasions (as needed, for example, to test the hypothesis of measurement equivalence). Overidentified models (which have more restrictions than are necessary to identify the model parameters) place restrictions on the hypothesized form of Σ, which may be used to test the goodness of fit of the model to the data using the likelihood test statistic. Differences in chi-square between *nested* models (models that have the same specification, with additional restrictions in one model) may be used to test the null hypothesis that the restrictions (e.g., constrained equal factor loadings) are true in the population.

In multiple group, longitudinal factor analysis, it is necessary to estimate factor models using covariance metric and sample covariance matrices

rather than to analyze separately standardized correlation matrices. Standardization could obscure invariant factor structures because of group differences in observed variances (Jöreskog, 1971), and would not allow evaluation of longitudinal changes in factor variances. To estimate raw score factor pattern weights and factor variances, one must identify the metric of the factors by fixing a single regression in each column of Λ to a constant (conveniently, 1.0), and then interpret results while considering the metric of latent and observed variables. The analyses reported here do so. Nevertheless, as standardized factor loadings (etc.) are easier to interpret, we provide parameter estimates that have been rescaled to a quasi-standardized metric, using a SAS PROC MATRIX program for scaling longitudinal factor analyses.[2] This rescaling preserves longitudinal constraints on parameter estimates but returns scaled values for factor loadings that are similar to standardized factor loadings. We also report maximum likelihood estimates and standard errors for certain models so that the reader may evaluate (a) a null hypothesis that each parameter is equal to zero, or (b) that group differences in unconstrained parameters are statistically reliable. In general, parameters that exceed their standard errors by a ratio of 2:1 are reliably different from zero at a 5% (per comparison) alpha level.

Results

The longitudinal models we estimate are designed to test the properties of the second-order general intelligence factor (g) from the PMA identified by Thurstone and Thurstone (1941). A first step was to determine that the g factor was an adequate representation of the covariance structure of the five PMA subtests. Bechtoldt (1974) and Corballis and Traub (1970) worked with a two-factor representation of the PMA subtests, although Bechtoldt's work included an additional memory variable that was not included in the 1948 PMA, and Corballis and Traub's two-factor model appeared to produce a very weak second factor. Nevertheless, we considered it necessary to evaluate the sufficiency of the g factor model before proceeding to longitudinal analysis. To do so, we used an exploratory factor analysis of all first-occasion cross-sectional data from the SLS (N = 2,202) to estimate an unrestricted maximum likelihood factor solution. The results for the one-factor model clearly indicated that the g factor sufficiently accounted for the covariance structure, $\chi^2(5, N = 2,202) = 6.18$, $p < .25$; Tucker-Lewis reliability = .997.

Longitudinal Model, Sample 1

Prior to analyzing the multiple age groups, we first analyzed the longitudinal factor model for the entire Sample 1. This analysis permitted us to evaluate the structural model prior to engaging in the more complex multiple group models reported later in the article. The basic occasion-specific model is depicted in Figure 1. The g factor was specified at each longitudinal occasion. The metric of g was defined by fixing the loading of Reasoning on g to 1.0. The remaining four factor loadings at each occasion were freely estimated, but were constrained to be equal across longitudinal occasions. By design, the loadings of all of the other variables (e.g., Verbal Meaning at Time 3 on g at Time 1) were fixed at 0. The factor covariance matrix was freely estimated, and the residual covariance matrix was specified as a diagonal matrix of unique variances.

We hypothesized in advance that this model would not fit the data because of the diagonal specification for the residual covariance matrix. It is well-known that longitudinal factor models of the type we are working with are likely to require what has been termed *autocorrelated* residuals (Sörbom, 1975; Wiley & Wiley, 1970). That is, given that it is likely that the occasion-specific factors will not account for all the reliable variance in the observed variables, then it is plausible to expect that the residuals (specific components) for an observed variable will correlate over time. In other words, we expected a residual covariance between the residual for Verbal Meaning at Time 1 and the Verbal Meaning residual at Time 2, a residual covariance between the Time 1 Space residual and the Time 2 Space residual, and so on. This residual pattern was especially likely, given that we are estimating a second-order g factor, as in this case the residual will include variance in the primary ability not accounted for by g. In fact, one would expect from the literature on abilities that the communalities for variables like Space and Number determined by g would be relatively small.

The initial model, denoted 0_1, specifying a diagonal matrix of unique variances provided an exceptionally poor fit to the data (see Table 2). The poor fit was underscored by the fact that the estimated factor covariances were greater than the corresponding factor variances (which implies the logical absurdity of correlations greater than 1). We therefore estimated Model 0_2, specifying autocorrelated residuals in the residual covariance matrix. The improvement in fit was substantial, change in $\chi^2(15, N = 162) = 898.64$, $p < .001$. Indeed, the overall chi-square test statistic was no longer significant, and the normed fit index was .96, indicating that nearly all the covariance in the sample data matrix was accounted for by the model.

At this point, our interest shifted to testing hypotheses regarding cross-occasion invariance in the parameter matrices. The principal hypothesis of interest with respect to measurement equivalence involved the invariance of the raw-score factor pattern weights (factor loadings) in Λ. Model 0_3 relaxed the constraint that the factor pattern weights be equal across occasions. The difference in fit was nonsignificant, indicating that the hypothesis of equal weights could not be rejected.

Given invariant factor pattern weights, it was meaningful to ask whether the factor variances were stationary over time, indicating consistency in the magnitude of individual differences on g. Model 0_4 tested this hypothesis by constraining the diagonal elements of the factor covariance matrix to be equal across longitudinal occasions. This hypothesis was rejected (see Table 2). Thus we concluded that there were changes in the magnitude of individual differences over occasions. We were also able to reject the null hypothesis that the factor covariances were equal (see Model 0_5 of Table 2).

Next, our attention turned to the parameters in the residual covariance matrix. Our first hypothesis was that the residual covariances could be constrained equal over occasions. This hypothesis, if tenable, would suggest a high degree of stability of individual differences in the ability-specific residual components. As can be seen in Table 2, Model 0_6, imposing the equality constraints on the residual covariances, did not fit worse than the Model 0_3, indicating that the hypothesis of equal covariances

[2] Briefly, the scaling is accomplished by pooling estimated latent variances and estimated observed variances to obtain scaling matrices. Pooling is done over multiple groups, as in Jöreskog (1971), and also over longitudinal occasions. A set of scaling equations and a listing of the scaling program is available from the first author on request.

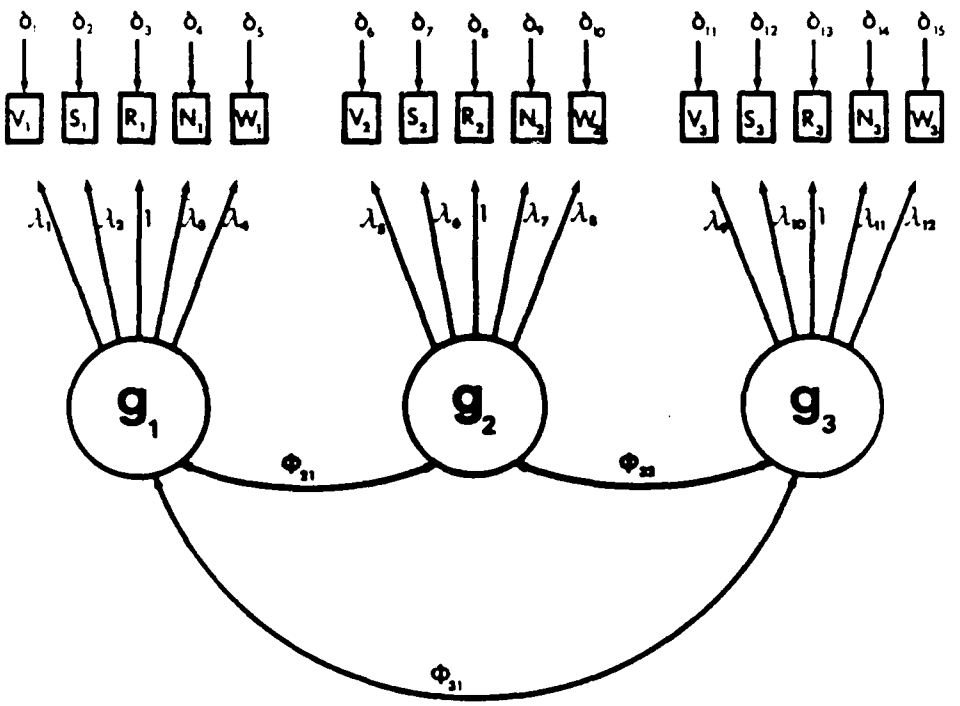


Figure 1. Initial longitudinal factor model specifying general intelligence factor (g) at each of three longitudinal occasions. (Subsequent models include covariances among corresponding residuals [e.g., δ_1, δ_6, δ_{11}] over time.)

could not be rejected. Finally, we tested the hypothesis of longitudinal invariance in the residual variances. This hypothesis stipulates that longitudinal changes in the variances of the observed variables could be attributed to changes in g factor variance alone. This model, labeled 0_7 in Table 2, was rejected as an equivalent representation to Model 0_6. We concluded that there were occasion-specific differences in the unique variances as well as in the factor variances.

The factor loadings their associated standard errors of the accepted model (0_6) are given in Table 3. All factor loadings are significant, but the rescaled factor loadings for Verbal Meaning and Reasoning are clearly larger than the rest. This pattern is consistent with the factor analytic literature on second-order ability factors (e.g., Horn, 1978), and parallels the findings of Thurstone and Thurstone (1941).

This pattern is also reflected in the standardized residual variances, where the smallest residuals (largest communalities) are associated with Verbal Meaning and Reasoning. Note also the

Table 2
Goodness-of-Fit Statistics for Alternative Longitudinal Models

Model	χ^2	*df*	*p*	ρ^a	Comparison	$\Delta\chi^2$	Δdf	*p*	$\Delta\rho$
$0_1(\Lambda_t =^b \text{diag } \Phi^e)$	985.84	95	.000	.574	—	—	—	—	—
$0_2(\Lambda_t =^b \text{cov } \Phi^d)$	87.20	80	.27	.962	0_1–0_2	898.64	15	< .001	.388
$0_3(\Lambda_t \neq)$	82.98	72	.17	.964	0_1–0_3	4.22	8	*ns*	.002
$0_4(\Lambda_t =^b, \text{diag } \Phi_t =^e)$	112.90	82	.013	.951	0_4–0_2	25.70	2	< .001	.011
$0_5(\Lambda_t = \Phi_t =^f)$	121.78	84	.005	.947	0_5–0_4	8.88	2	< .05	.004
$0_6(\Lambda_t =^b, \text{cov } \Theta =^g)$	97.16	90	.28	.958	0_6–0_2	9.96	10	*ns*	.004
$0_7(\Lambda_t = \Theta_t =^h)$	129.21	100	.026	.944	0_7–0_6	32.05	10	< .05	.014

[a] Bentler–Bonett normed fit index.
[b] Indicates nonzero factor pattern weights in Λ constrained to be equal over time (t).
[c] Indicates the residuals in Θ specified as uncorrelated (see text).
[d] Indicates autocorrelated residuals in Θ. This specification was continued in Models 0_3–0_7, as well.
[e] Indicates factor variances in Φ constrained to be equal over time.
[f] Indicates factor covariances constrained equal, and factor variances constrained equal over time.
[g] Indicates covariances among residuals constrained equal over time.
[h] Indicates residual variances constrained equal over time, and residual covariances constrained equal over time.

Table 3
Factor Loadings and Residual Variances for the Longitudinal Factor Model (0_6)

Test	Factor loadings		Residual variances[a]		
	LISREL estimates[b]	Rescaled loadings	Time 1	Time 2	Time 3
Verbal Meaning	1.540 (0.100)	.838	.318	.348	.240
Space	0.994 (0.109)	.556	.751	.666	.652
Reasoning	1.00[c] (—)	.878	.269	.274	.162
Number	0.928 (0.108)	.518	.760	.763	.674
Word Fluency	1.108 (0.133)	.520	.774	.735	.682

[a] Calculated as the proportion of residual variance (estimated) to total variance (estimated); 1 − (residual variance) = the communality.
[b] Standard errors in parentheses.
[c] Fixed parameter.

longitudinal decreases in residual variances for all variables, suggesting that the communalities of the primary ability variables determined by *g* increase over time. The high degree of stability in individual differences is reflected in the high factor covariances, which are provided in Table 4. Standardized, these covariances reflect correlations of greater than .9 between *g* at each longitudinal occasion. Clearly, there is not much change in the relative ordering of individuals on general intelligence over the 14-year period.

The results of this model were successfully cross validated in Sample 2. Rather than report these results, we move immediately to discussion of the multiple group analysis.

Multiple Group Analysis

The analyses in Samples 1 and 2 suggest almost perfect stability of individual differences in intelligence, both at the *g* factor and test-specific component levels. These analyses combined individuals spanning the adult life span, however, and it was possible that the wide age range served to maximize the apparent stability of individual differences. In particular, it was possible that differential change in the late-middle-age/old-age ranges was obscured by the high degree of stability across most of the adult life span. The multiple group analyses were designed to examine the stability of individual differences in more homogeneous age ranges. They also afforded us the opportunity of looking at age group differences in the factor analysis parameters. One might expect that there would be a greater opportunity for age group differences in factor loadings—given the age ranges spanned by our groups—than for longitudinal age changes.

We began by testing the equality of the observed covariance matrices across the three age groups. Box's test suggested non-homogeneous covariance matrices, $M = 402.77$, $F(240, \infty) = 1.59$, $p < .0001$. This result made it likely that there indeed were group differences in some of the factor analytic parameters.

Table 4
Factor Covariance Matrix (and Correlations) for the Longitudinal Factor Model (0_6)

Factor	g_1	g_2	g_3
g_1	28.624 (4.137)	0.945	0.917
g_2	27.723 (3.983)	30.062 (4.338)	0.972
g_3	31.776 (4.531)	34.528 (4.787)	41.938 (5.728)

Note. g_1 is the general factor at Time 1, g_2 is the general factor at Time 2, g_3 is the general factor at Time 3. Standard errors in parentheses. Values above the diagonal are standardized factor correlations.

The longitudinal factor model investigated in Sample 1 was used in the multiple group analyses. However, rather than presume the equivalence of residual covariances (as in Model 0_6 above) we chose to begin with these parameters unconstrained. Our rationale was that group differences in the residual covariance structure might have been obscured in the single sample analysis. Rather than presume the constraints, we chose to evaluate them anew in the multiple group model.

Our basic model, then, posited the specification of Model 0_5 of the Sample 1 analyses: an occasion-specific *g* factor (with no longitudinal constraints on the factor loadings), a freely estimated factor covariance matrix, and a residual covariance matrix with free unique variances and autocorrelated residual covariances. This model was specified in each of the three age groups, with no additional constraints on the parameters across the groups. The model was therefore equivalent to running the longitudinal factor model separately in the three groups.

As can be seen from the first entry in Table 5, this model, denoted M_1, provided a relatively good fit to the data, allowing us to conclude that it was a reasonable representation of the covariance matrices in each group. We therefore proceeded to test for invariance in the *g* factor loadings. Separate tests of the equality of the factor loadings across age groups (Model M_2) and longitudinally across occasions (Model M_3) did not fit worse than the model with no constraints on the factor loadings (see Table 5). For both tests, the combined change in chi-square was actually just less than the change in degrees of freedom, $\chi^2(32, N = 412) = 29.82$, *ns*. We therefore concluded that the *g* factor loadings demonstrated complete age equivalence—being invariant both longitudinally and between age groups.

Our next set of models examined invariance in the factor covariance matrix. Model M_4, requiring age group equivalence in the factor covariances matrix (both variances and covariances), significantly degraded the fit to the data, requiring rejection of the null hypothesis of age group equivalence. We next tested a less restrictive model, positing group equivalence in factor variances but not in covariances. This model (M_5) was also rejected. Finally, Model M_6, placing no group constraints on the variances but positing longitudinal equality of variances within each group, was also rejected by the data (see Table 5). We should note that none of these models greatly degraded the fit, as judged by the normed fit index change of .01 or less (see Bentler & Bonett,

Table 5
Goodness-of-Fit Statistics for Models With Multiple Groups

Model	χ^2	*df*	*p*	ρ^a	Comparison	$\Delta\chi^2$	Δdf	*p*	$\Delta\rho$
M_1 (all free)[b]	257.85	216	.027	.951	—	—	—	—	—
$M_2(\Lambda_g = \Lambda)$[c]	284.24	240	.026	.946	M_2–M_1	26.39	24	*ns*	.005
$M_3(\Lambda_i = \Phi)$[d]	287.68	248	.042	.945	M_3–M_2	3.44	8	*ns*	.001
$M_4(\phi_g = \phi)$[e]	329.65	260	.002	.937	M_4–M_3	41.97	12	< .01	.008
$M_5(\text{var}\phi_g = \phi)$[f]	310.68	254	.004	.941	M_5–M_3	23.00	6	< .01	.004
$M_6(\text{var}\phi_i = \phi)$[g]	301.28	254	.022	.943	M_6–M_3	14.00	6	< .05	.002
$M_7(\Theta_g = \theta)$[h]	458.85	308	.000	.913	M_7–M_3	171.17	60	< .001	.032
$M_8(\text{cov}\Theta_i = \gamma)$[i]	331.77	278	.015	.937	M_8–M_3	44.09	30	< .05	.008

[a] Bentler-Bonett normed fit index.
[b] Indicates no between-groups equality constraints among parameters.
[c] Indicates factor loadings constrained equal between groups.
[d] Indicates factor loadings constrained equal between groups (as in M_2) *and* constrained equal over time (this specification maintained in Models M_4–M_8).
[e] Indicates factor covariance matrices constrained equal between groups.
[f] Indicates factor variances constrained equal over groups.
[g] Indicates factor variances constrained equal over time in each of the groups.
[h] Indicates entire residual covariance matrix constrained equal over groups.
[i] Indicates residual covariances for test-specific components constrained equal over time.

1980). Nevertheless, the loss of fit, judged from the likelihood ratio chi-square test, was significant. These results indicated that the factor covariance matrices should neither be taken to be stationary over time nor equivalent across age groups.

Finally, we pursued the residual covariance structure to assess the stability of the residual variances and covariances across time. A preliminary model, M_7, specified group invariance in all three parameter matrices (Λ, Φ, and Θ). Compared to model M_3, this model tests the age group equivalence of the residual covariance matrix. The hypothesis was convincingly rejected. Our next step was to evaluate the plausibility of a model constraining the residual covariances to be equal between different measurement occasions (as was the case for Model 0_6 in the single sample analysis). Model M_8 placed these constraints on the residuals. The loss of fit was marginally significant at the 95% confidence level. We concluded that the model specifying equal covariances had missed the mark, but not by much. Thus, unlike Model 0_6, we could not treat the residual covariances as invariant over longitudinal occasions in the multiple group analysis. Apparently, both the residual variances and covariances differed by group and over longitudinal occasions, although the loss of fit due to group constraints was clearly much greater than the loss due to fitting invariant residual covariances over longitudinal occasions in each of the groups separately.

An alternative method for approaching stability in the residual covariances is by specification of a model positing both occasion-specific and test-specific factors (e.g., Jöreskog & Sörbom, 1977). Figure 2 depicts the factor pattern matrix (Λ) associated with a combined occasion-specific and test-specific factor model for these data. A given variable loads both on the general factor and its own test-specific factor (i.e., a Verbal Meaning factor, a Space factor, and so on). This parameterization of the residual covariances is plausible if one argues for a special relation among the residuals over time—a first order autoregressive structure (see Jöreskog & Sörbom, 1977). Addition of test-specific factors places no additional restrictions on the residual covariances, given that there are only three occasions of measurement (with more occasions, specification that the residual covariances form a single common factor may not fit the residual covariance structure). The advantage of the test-specific factor representation is that it enables one to separately estimate components of variance associated with *g*, stable variance in the primary ability, and a residual consisting of unstable variance plus measurement error (see Hertzog, in press).

We reestimated model M_3 (invariant factor loadings only) with test-specific factors. The parameter estimates and standard errors are provided in Tables 6 and 7. Given the fact that the hypothesis of invariant *g* factor loadings had been found plausible, we were entitled to assume measurement equivalence and to evaluate the remaining parameter estimates with respect to the issue of stability and change in intelligence. Several points of interest regarding the stability of individual differences emerged. First, the factor covariances were again extremely high, indicating a great degree of stability in individual differences in *g* over the 14-year interval for all three age groups. Standardized, these factor correlations are approximately .9 (or greater) for all groups (see Table 7).

Table 8 summarizes the stability of individual differences by reporting the correlations, r^2, and the estimated autoregressive coefficients predicting *g* from the previous longitudinal occasion. As can be seen from Table 9, the r^2 is larger for g_2 to g_3 in all groups, accounting for 92% of the variance in g_3 in both the middle-aged and old groups. The predominance of stability is underscored by the regression coefficients reported in Table 9. As suggested by Kessler and Greenberg (1981), we have expressed the raw-score slope coefficients in terms of the stability and, as given in the last column of Table 9, the regression of the change scores on initial scores (e.g., the regression of g_3–g_2 on g_2). This latter coefficient, if negative, suggests regression to the mean; if positive, it suggests increasing differences between individuals that covary with initial differences. Table 6 shows that the raw-score slopes were very near 1.0 (suggesting high stability) and that the change slopes were near zero (suggesting little change variance predictable from initial scores). In both the middle-

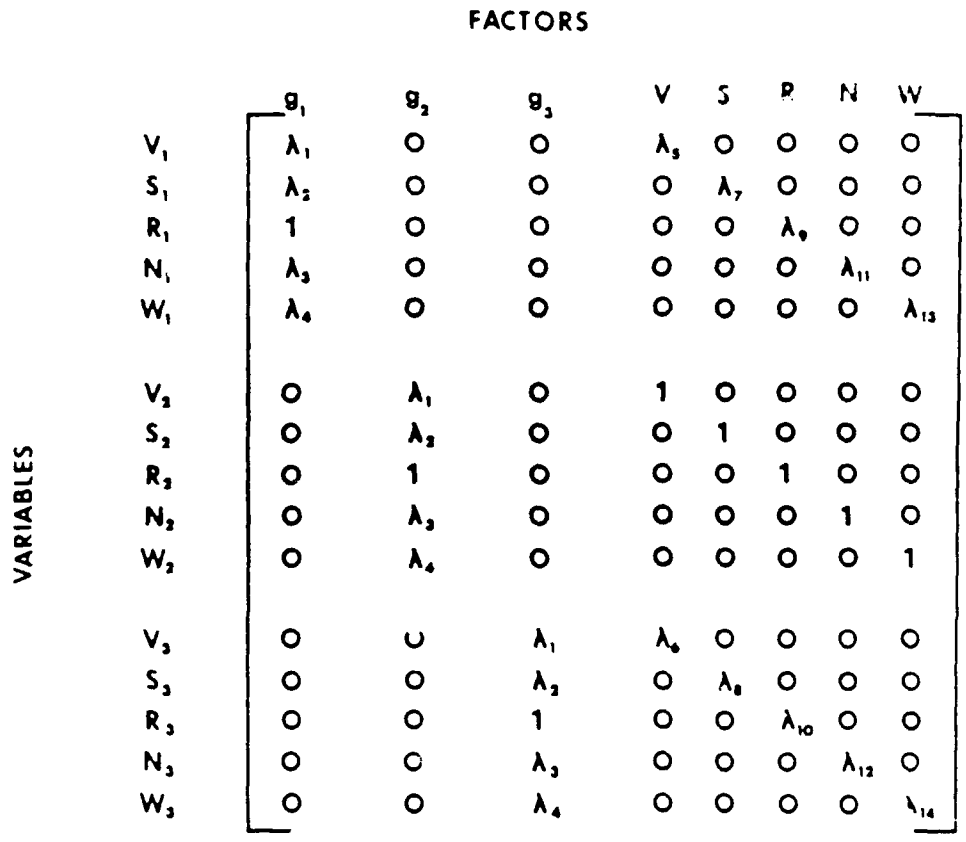


Figure 2 Factor pattern matrix for model including occasion-specific and test-specific factors (0's and 1's are fixed parameters, λ's are estimated by the model.)

aged and old groups, the change slopes were slightly negative for g_1 and g_2, suggesting slight regression to the mean, and slightly positive from g_2 to g_3, suggesting some egression from the mean (the rich getting richer, the poor poorer, as it were). In the young group, the stabilities were lower, albeit still impressively large, and the regression to the mean was consistent across time intervals.

The patterns of stability and change identified in the regression coefficients were mirrored in the factor variances, which exhibited different patterns of change across each of the groups. Factor variances decreased in the young group, but showed reliable increases from the second to the third occasion of measurement in both the middle-aged and old groups. This increase in *g* variance was consistent with the regression from the mean suggested from the regression coefficients. The decreases in variance and the regression to the mean pattern in the young group may reflect the mild ceiling effects on Verbal Meaning and Reasoning that we have observed in the youngest age groups in the SLS longitudinal samples.

Third, factor variances varied in magnitude between the age groups. The older group was generally more heterogeneous (had greater individual differences in *g*) than were the young and middle-aged groups. Taken together, these results suggested that although there was significant stability of individual differences in all age groups, the old group showed an interesting pattern of (a) greater variability in *g* at initial measurement and (b) increasing variability over time.[3]

An alternative way of looking at stability is the decomposition of variance in the model including both occasion-specific and test-specific factors. As can be seen in Table 9, the preponderance of *g* variance at the second and third occasions of measurement is stable variance predicted by individual differences at the prior measurement occasion. Given that we were studying the second-order *g* factor, it is relevant to ask about the stability of the residual components, reflecting the five primary ability factors from the PMA. Table 9 reports the decomposition of variance on each of the 15 observed variables for each group into proportions of (a) *g*-related variance, (b) stable test-specific variance, and (c) residual variance. The *g*-related variance components are actually the communalities of the observed variables with respect to the *g*

[3] One concern we had was that the patterns of factor variances might be due to the different age span for the oldest group (see Table 1). We therefore reanalyzed the data, using only the two oldest cohorts in Samples 1 and 2 to form a smaller old group. The redefinition of the old group did not eliminate the higher variances in *g* for the old, but did attenuate the longitudinal increases in variance. This analysis is discussed in more detail in the second article in this series (Hertzog & Schaie, 1986).

Table 6
Factor Loadings for Model With Occasion-Specific (g) and Test-Specific Factors

Variable	g[a]	g[ab]	Test (Young)[c]	Test (Middle aged)[c]	Test (Old)[c]
V_1	1.659 (.098)	.767	1.032 (.129)	0.921 (.122)	0.650 (.193)
S_1	0.948 (.087)	.438	1.001 (.084)	0.908 (.107)	1.136 (.208)
R_1	1.000* —	.777	0.752 (.174)	1.120 (.151)	0.708 (.199)
N_1	1.463 (.106)	.588	1.005 (.086)	0.962 (.058)	0.935 (.084)
W_1	1.340 (.118)	.485	0.667 (.102)	1.049 (.102)	1.046 (.104)
V_2	1.659 (.098)	.767	1.000* —	1.000* —	1.000* —
S_2	0.948 (.087)	.438	1.000* —	1.000* —	1.000* —
R_2	1.000* —	.777	1.000* —	1.000* —	1.000* —
N_2	1.463 (.106)	.588	1.000* —	1.000* —	1.000* —
W_2	1.340 (.118)	.485	1.000* —	1.000* —	1.000* —
V_3	1.659 (.098)	.767	0.971 (.120)	0.820 (.117)	1.042 (.323)
S_3	0.948 (.087)	.438	0.965 (.089)	0.770 (.095)	1.130 (.211)
R_3	1.000* —	.777	0.920 (.208)	1.006 (.133)	0.740 (.196)
N_3	1.463 (.106)	.588	0.970 (.080)	0.868 (.053)	0.786 (.074)
W_3	1.340 (.118)	.485	0.988 (.126)	0.925 (.086)	0.928 (.092)

Note Standard errors are in parentheses. Asterisks denote fixed parameters. Subscripts on variables indicate longitudinal occasion (1 = Time 1, 2 = Time 2, 3 = Time 3). V = Verbal Meaning; S = Space; R = Reasoning, N = Number; W = Word Fluency
[a] Factor loadings for occasion-specific general factor (g). Estimates were constrained equal across the 3 longitudinal occasions.
[b] Rescaled general factor loadings.
[c] Test-specific factor loadings for each age group.

factor. The variance associated with the test-specific factor represents stable variance across occasions specific to the primary ability. The residual variance represents a combination of measurement error variance and unstable specific variance (the two components cannot be disentangled in this analysis). There are several points of interest in Table 9. First, the communalities of the g factor increased substantially in the old group relative to the young and middle-aged groups (and showed a tendency to increase over time longitudinally as well). Thus g determines more of the variance of the observed measures in the old than in the young. Second, those variables with the lowest communalities for g (Space, Number, Word Fluency) show very high levels of stability in the primary ability (test-specific) domain. For example, although only about 14% of the young group's variance of Space at Time 1 is determined by g, 72% of Space's Time 1 variance is determined by the Space test-specific factor in the young group. This indicates substantial stability in both the g and test-specific domains. Proportions of stable test-specific variance to total g-adjusted variance are given in the right-hand column of Table 9. Considering that these proportions are contaminated by measurement error, the proportion of stable variance in the primary ability measures independent of g is indeed impressive. Finally, the unique variances show some evidence of change in the primary abilities, but in many cases the proportions of unique variance are close to what would be expected to be the magnitude of error variance, given the reliabilities of the measures reported by Thurstone and Thurstone (1949).

Table 7
Factor Covariance Matrices for Occasion-Specific Factors in Each Age Group

Factor	g_1	g_2	g_3
	Young		
g_1	15.048 (2.868)	0.887	0.930
g_2	11.896 (2.409)	11.959 (2.421)	0.933
g_3	11.951 (2.365)	10.690 (2.179)	10.970 (2.257)
	Middle aged		
g_1	16.797 (2.691)	0.927	0.960
g_2	16.204 (2.549)	16.761 (2.652)	0.959
g_3	16.786 (2.607)	16.760 (2.591)	18.204 (2.798)
	Old		
g_1	23.546 (3.595)	0.944	0.885
g_2	22.405 (3.427)	23.941 (3.713)	0.959
g_3	23.442 (3.598)	25.589 (3.814)	29.769 (4.335)

Note Standard errors are in parentheses. Values above the diagonal are factor correlations, standardized independently in each age group.

Table 8
Correlations and Regression Coefficients Indicating Stability of Individual Differences in g

Group	r[a]	r^2	$1-r^2$	b[b]	b_{xx}[c]
Young					
g_1, g_2	.887	.787	.213	0.791	−0.209
g_2, g_3	.933	.870	.130	0.894	−0.106
Middle aged					
g_1, g_2	.927	.859	.141	0.965	−0.035
g_2, g_3	.959	.920	.080	1.000	0.000
Old					
g_1, g_2	.944	.891	.109	0.952	−0.048
g_2, g_3	.959	.920	.080	1.069	0.069

Note Stabilities are shown for 7-year intervals between adjacent longitudinal occasions.
[a] Simple correlation of scores for adjacent occasions.
[b] Simple regression of later occasion on earlier occasion (unstandardized)
[c] Regression of change score on earlier occasion (unstandardized).

Table 9
Estimated Variance Components From Final Multiple Groups Model

Variable	$\hat{\sigma}^2$	g^a	Test-specific[b]	Unique[c]	Stable (test)[d]
			Young		
V_1	76.286	.543	.333	.124	.729
S_1	98.688	.137	.724	.139	.839
R_1	27.518	.547	.186	.268	.410
N_1	114.107	.282	.591	.127	.823
W_1	136.097	.199	.332	.470	.414
V_2	62.008	.531	.385	.084	.821
S_2	103.512	.104	.689	.208	.768
R_2	27.832	.430	.325	.246	.569
N_2	115.828	.221	.576	.203	.739
W_2	148.662	.144	.683	.173	.798
V_3	63.587	.475	.354	.171	.674
S_3	104.872	.094	.634	.274	.698
R_3	24.025	.457	.318	.225	.586
N_3	96.291	.244	.652	.104	.862
W_3	159.879	.123	.620	.257	.713
			Middle aged		
V_1	77.263	.506	.273	.128	.681
S_1	97.719	.153	.468	.369	.559
R_1	32.471	.517	.299	.184	.448
N_1	120.387	.299	.589	.112	.840
W_1	154.861	.195	.502	.304	.623
V_2	81.848	.564	.304	.133	.696
S_2	82.420	.183	.680	.137	.832
R_2	29.076	.576	.266	.157	.629
N_2	127.997	.280	.599	.121	.832
W_2	125.227	.240	.564	.196	.742
V_3	83.260	.617	.206	.178	.536
S_3	88.797	.191	.387	.422	.476
R_3	30.568	.596	.256	.148	.634
N_3	109.363	.356	.528	.116	.820
W_3	119.517	.273	.505	.221	.696
			Old		
V_1	102.167	.634	.087	.278	.238
S_1	83.784	.253	.348	.400	.465
R_1	34.374	.685	.105	.210	.333
N_1	119.696	.421	.441	.138	.762
W_1	163.650	.255	.516	.226	.695
V_2	115.064	.573	.184	.243	.431
S_2	77.426	.278	.292	.431	.404
R_2	36.005	.665	.201	.134	.600
N_2	129.347	.396	.466	.138	.772
W_2	152.787	.281	.505	.214	.702
V_3	126.724	.647	.182	.172	.514
S_3	74.825	.356	.385	.258	.599
R_3	38.027	.783	.104	.113	.479
N_3	119.211	.534	.313	.153	.672
W_3	151.523	.353	.438	.208	.678

Note $\hat{\sigma}^2$ = estimated variance of observed variable. V = verbal meaning, S = space, R = reasoning, N = number, W = word fluency. Subscripts on variables indicate longitudinal occasion (1 = Time 1, 2 = Time 2, 3 = Time 3).
[a] Proportion of variance due to *g*.
[b] Proportion of variance due to test-specific factor.
[c] Proportion of variance unique to the observed variable. The sum of the three proportions (*g*-related, test-specific, unique) is 1.0.
[d] Proportion of variance *not* determined by *g* that is determined by the test-specific factor.

Discussion

The results of the present study present a relatively coherent picture—one of measurement equivalence and stability in psychometric intelligence, as measured by the Thurstones's 1948 Primary Mental Abilities test, in adulthood. We found that it was highly plausible to model the factor loadings of a general intelligence factor as being invariant, both longitudinally and across multiple age groups. We also found a high degree of stability of individual differences across the adult life span.

The finding of invariance in the *g* factor loadings is important relative to the suggestion in the literature that the fundamental measurement properties of the psychometric tests change over the life span (e.g., Baltes & Nesselroade, 1970; Demming & Pressey, 1957; Schaie, 1977). As shown by Meredith (1964), under selection of subpopulations from a population for which an isomorphic common factor model holds, the multiple subpopulations will have an invariant unstandardized factor pattern matrix. Meredith's work implies that one must reject the hypothesis of metric invariance before one is justified in concluding that the groups have qualitatively different factor structures. One cannot argue for qualitative group differences in measurement properties if the hypothesis of metric invariance cannot be rejected. In contrast, we found the hypothesis of metric invariance to be strongly supported by our data. Our results therefore suggest that, whatever the faults inherent in the constructs of psychometric intelligence, measures of psychometric intelligence seem to be measuring basically isomorphic constructs with similar measurement properties at different age levels.

One could still, of course, argue that the constructs measured by psychometric intelligence are of limited utility in predicting intelligent behaviors in adults (e.g., Sternberg, 1985). Nevertheless, our findings do not support the notion that psychometric testing of abilities in older populations is invalid because one is measuring qualitatively different constructs with unstable measures. Our conclusion must be qualified by the fact that our assessment of factorial invariance is specific to the second-order *g* factor. We cannot assess the invariance of the primary ability factor loadings from our data. We therefore cannot rule out the possibility of nonequivalent measurement properties at the primary ability level, although, given the stability indicated by the test-specific factors, the likelihood of measurement equivalence in the primary ability factors seems quite high. Data we recently collected on an expanded ability battery as part of the 1984 SLS assessment should help us address the measurement equivalence issue at the primary ability factor level.

The finding of factorial invariance is relevant to the factor analytic literature suggesting de-differentiation of ability factors in old age (Reinert, 1970). The de-differentiation argument states that ability factors coalesce, or collapse, toward a general intelligence factor in older groups. The early literature on this phenomenon was plagued by methodological inadequacies (Cunningham, 1978; Reinert, 1970; Schaie & Hertzog, 1985). Recent comparative factor analysis work by Cunningham (1980, 1981), using confirmatory factor analysis methods, suggests that there is little evidence for gross collapse of the factor space—the same number of factors are needed to model ability variables in old groups, and the loading patterns are highly similar. Our results are consistent with Cunningham's findings in suggesting invari-

ance in the raw-score regressions of variables on ability factors, both across age groups and longitudinally within age groups (see also Cunningham & Birren, 1980).

Cunningham (1980, 1981) reported evidence for a mild form of de-differentiation—that is, increased factor correlations in the older groups. Our finding of increased communalities for *g* in the old group is also consistent with this mild form of de-differentiation. To clarify the relation, we report in Table 10 correlations among the primary abilities obtained by a confirmatory factor analysis specifying test-specific factors. As can be seen in Table 10, there is a pronounced tendency for factor correlations to be higher in the old group. Crude indexes of this tendency are the average correlations of .36 for the young group, .39 for the middle-aged group, and .54 for the old group. Nevertheless, it must be emphasized that the primary thrust of the de-differentiation argument—qualitative change in the nature of ability factors—is neither supported by Cunningham's findings nor by our own.

The age-related measurement equivalence in the PMA allows us to make unambiguous interpretation of the stability of individual differences in *g* over time. Clearly, individual differences in general intelligence are highly stable across 14-year longitudinal epochs for three age groups (spanning most of the adult age range). The stability coefficients indicated that approximately 90% of the *g* variance in the middle-aged and old groups was consistent between adjacent 7-year testing intervals. There is, then, little indication in these data of any substantial degree of variability in developmental trajectories in *g*. Moreover, the stability of individual differences in the PMA ability-specific components in our longitudinal model suggest a high degree of stability in individual differences on the primary abilities as well.

Although these results clearly limit the degree to which one could argue for a substantial degree of interindividual differences in intraindividual change in psychometric intelligence in adulthood, it would be overstating the case to argue that these data demonstrate a *lack* of variability in change functions across the adult life span. For one thing, it is well-known that the longitudinal samples of the SLS are influenced by a substantial degree of experimental mortality (Schaie, Labouvie, & Barrett, 1973), causing the participants in the 14-year studies to be relatively select with respect to ability levels. It is highly likely, given the relatively long 7-year retest interval and the nature of the sampling procedures, that individuals in terminal decline or suffering differential loss of abilities due to severe illness will have dropped out of the longitudinal sample (Hertzog, Schaie, & Gribbin, 1978). The high degree of stability we observed in this study may be specific to more select, healthy subpopulations of adults and may not generalize to the population at large. Moreover, our sample size was sufficiently small that we were forced to pool over relatively large age ranges to form our age groups. Such a procedure maximizes individual differences at the initial measurement occasion and may have obscured some degree of heterogeneity in developmental trends. We note, however, that the estimates of stability did not differ greatly between the Sample 1 analysis and the age-partitioned multiple group analysis that reduced individual differences produced by wide age spans.

Of course, as McCall (1981) pointed out, even stabilities of .9 allow for a greater degree of crossover of individual curves than might be expected by social scientists. At the individual level, it is still possible that a given individual will buck the tide, and exhibit less change in *g* than his or her same-age peers. There may also be more variability in the primary abilities than in the higher order intelligence factor. One can see in Table 7 that the test-specific stabilities were in some cases smaller than the stabilities for *g* in the same age interval. In the old group, for example, the stability of the Space test-specific factor seems to be smaller than the stability observed for Space in the young and middle aged, even though the stability of individual differences in *g* is, if anything, greater in the old group. This result may indicate slightly more variability in the patterns for the Spatial Orientation ability tapped by the Space test (see McGee, 1979). These data are not optimally suited for assessing primary ability-specific change, however, because unreliability due to measurement error cannot be separated from instability in the ability in the analysis we have reported. In any case, we must be careful to emphasize that there is considerably more consistency than inconsistency in age changes in all age groups, and for all PMA subtests. Finally, we cannot rule out the possibility that individual differences in change (that matter, changes in factor loadings), occur in older ages (beyond 80) not represented in this study.

Table 10
Primary Ability Factor Correlations for the Three Age Groups

Variable	Verbal Meaning	Space	Reasoning	Number	Word Fluency
Young (M age = 37)					
Verbal Meaning	1				
Space	.115	1			
Reasoning	.559	.455	1		
Number	.390	.239	.489	1	
Word Fluency	.531	.034	.425	.334	1
Middle aged (M age = 49)					
Verbal Meaning	1				
Space	.296	1			
Reasoning	.711	.479	1		
Number	.419	.248	.441	1	
Word Fluency	.508	.039	.439	.308	1
Old (M age = 65)					
Verbal Meaning	1				
Space	.593	1			
Reasoning	.838	.650	1		
Number	.666	.528	.627	1	
Word Fluency	.557	.290	.202	.450	1

The invariance in the PMA *g* factor loadings and the stability of individual differences in intelligence contrasts sharply with patterns of mean age changes found in the SLS (e.g., Schaie, 1983; Schaie & Hertzog, 1983). Schaie has consistently found variation in mean patterns according to age, cohort, and time of measurement. Moreover, these mean changes have been found to vary in magnitude for different abilities. The difference in findings underscores the critical distinction between stability in means (i.e., on average, no age changes) and stability of individual differences. In normally distributed variables, stability of the means and stability of individual differences (as measured by

covariances) are statistically (and conceptually) independent. As one can see in the next article in this series (Hertzog & Schaie, 1986), we can observe stability of individual differences either when there are no mean age changes or when there are substantial mean changes over a given portion of the life span.

References

Baltes, P. B., Dittmann-Kohli, F., & Dixon, R. A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behaviors* (Vol. 6, pp. 34–76). New York: Academic Press.

Baltes, P. B., & Nesselroade, J. R. (1970). Multivariate longitudinal and cross-sectional sequences for analyzing ontogenetic and generational change: A methodological note. *Developmental Psychology, 1*, 162–168.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks-Cole.

Baltes, P. B., & Schaie, K. W. (1976). On the plasticity of intelligence in adulthood and old age: Where Horn and Donaldson fail. *American Psychologist, 31*, 720–725.

Bechtoldt, H. P. (1974). A confirmatory analysis of the factor stability hypothesis. *Psychometrika, 39*, 319–326.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 580–605). New York: Van Nostrand Reinhold.

Corballis, M. C. (1973). A factor model for analyzing change. *British Journal of Mathematical and Statistical Psychology, 26*, 90–97.

Corballis, M. C., & Traub, R. E. (1970). Longitudinal factor analysis. *Psychometrika, 35*, 79–93.

Cunningham, W. R. (1978). Principles for identifying structural differences: Some methodological issues related to comparative factor analysis. *Journal of Gerontology, 33*, 82–86.

Cunningham, W. R. (1980). Age comparative factor analysis of ability variables in adulthood and old age. *Intelligence, 4*, 133–149.

Cunningham, W. R. (1981). Ability factor structure differences in adulthood and old age. *Multivariate Behavioral Research, 16*, 3–22.

Cunningham, W. R., & Birren, J. E. (1980). Age changes in the factor structure of intellectual abilities in adulthood and old age. *Educational and Psychological Measurement, 40*, 271–290.

Demming, J. A., & Pressey, S. L. (1957). Tests "indigenous" to the adult and older years. *Journal of Consulting Psychology, 4*, 144–148.

Eckensberger, L. (1973). Methodological issues of cross-cultural research in developmental psychology. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 43–64). New York: Academic Press.

Hertzog, C. (1985). An individual differences perspective: Implications for cognitive research in gerontology. *Research on Aging, 7*, 7–45.

Hertzog, C. (in press). On the utility of structural regression models for developmental research. In P. B. Baltes, D. Featherman, & R. M. Lerner (Eds.), *Life-span development and behavior* (Vol. 8). Hillsdale, NJ: Erlbaum.

Hertzog, C., & Schaie, K. W. (1986). *Stability and change in adult intelligence 2: Simultaneous analysis of longitudinal means and covariance structures*. Unpublished manuscript, Georgia Institute of Technology, Atlanta.

Hertzog, C., Schaie, K. W., & Gribbin, K. (1978). Cardiovascular disease and changes in intellectual functioning from middle to old age. *Journal of Gerontology, 33*, 872–883.

Horn, J. L. (1978). Human ability systems. In P. B. Baltes (Ed.), *Life-span development and behavior* (Vol. 1, pp. 211–256). New York: Academic Press.

Horn, J. L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. *American Psychologist, 31*, 701–719.

Horn, J. L., & Donaldson, G. (1980). Cognitive development in adulthood. In O. G. Brim & J. Kagan (Eds.), *Constancy and change in human development* (pp. 445–529). Cambridge, MA: Harvard University Press.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical/statistical modeling (MASMOB) in research on aging. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.

Horn, J. L., McArdle, J. J., & Mason, R. (1984). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1*, 179–188.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D. S. Aigner & A. S. Goldberger (Eds.), *Latent variables in socio-economic models* (pp. 285–325). Amsterdam: North Holland.

Jöreskog, K. G., & Sörbom, D. (1980). *Simultaneous analysis of longitudinal data from several cohorts* (Research Rep. 80-5). Uppsala, Sweden: University of Uppsala, Department of Statistics.

Jöreskog, K. G., & Sorbom, D. (1981). *LISREL V users guide*. Chicago: National Educational Resources.

Kagan, J. (1980). Perspectives on continuity. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 26–74). Cambridge, MA: Harvard University Press.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change*. New York: Academic Press.

McCall, R. B. (1981). Nature-nurture and the two realms of development: A proposed integration with respect to mental development. *Child Development, 52*, 1–12.

McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin, 86*, 889–918.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177–185.

Nesselroade, J. R., & Labouvie, E. W. (1985). Experimental design in research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 35–60). New York: Van Nostrand Reinhold.

Reinert, G. (1970). Comparative factor analytic studies of intelligence throughout the human life span. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology: Research and theory* (pp. 467–484). New York: Academic Press.

Schaie, K. W. (1977). Quasi-experimental research designs in the psychology of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 39–58). New York: Van Nostrand Reinhold.

Schaie, K. W. (1979). The primary mental abilities in adulthood: An exploration in the development of psychometric intelligence. In P. B. Baltes & O. G. Brim (Eds.), *Life-span development and behavior* (Vol. 2, pp. 67–115). New York: Academic Press.

Schaie, K. W. (1983). The Seattle Longitudinal Study: A 21-year exploration of psychometric intelligence in adulthood. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 64–135). New York: Guilford Press.

Schaie, K. W., & Hertzog, C. (1982). Longitudinal methods. In B. B. Wolman (Ed.), *Handbook of developmental psychology* (pp. 91–115). Englewood Cliffs, NJ: Prentice-Hall.

Schaie, K. W., & Hertzog, C. (1983). Fourteen-year cohort-sequential analyses of adult intellectual development. *Developmental Psychology, 19*, 531–543.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 61–92). New York: Van Nostrand Reinhold.

Schaie, K. W., Labouvie, G. V., & Barrett, J. J. (1973). Selective attrition effects in a 14-year study of adult intelligence. *Journal of Gerontology, 28*, 328–334.

Schaie, K. W., Rosenthal, F., & Perlman, R. M. (1953). Differential mental deterioration of factorially "pure" functions in later maturity. *Journal of Gerontology, 8*, 191–196.

Schaie, K. W., & Willis, S. L. (1986). Can decline in adult intellectual functioning be reversed? *Developmental Psychology, 22*, 223–232.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 28*, 138–151.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of intelligence*. New York: Cambridge University Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence*. (Psychometric Monographs No. 2). Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1949). *Examiners manual: SRA Primary Mental Abilities test* (Form 11–17). Chicago: Science Research Associates.

Wiley, D. E., & Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review, 35*, 112–117.

Received May 3, 1985
Revision received October 21, 1985 ■

APPENDIX E

STABILITY AND CHANGE IN ADULT INTELLIGENCE: 2. SIMULTANEOUS ANALYSIS OF LONGITUDINAL MEANS AND COVARIANCE STRUCTURES

By

Christopher Hertzog
Georgia Institute of Technology

K. Warner Schaie
Pennsylvania State University

Psychology and Aging
1988 Vol 3, No 2, 122-130

0882-7974/88/$00.75

Stability and Change in Adult Intelligence: 2. Simultaneous Analysis of Longitudinal Means and Covariance Structures

Christopher Hertzog
School of Psychology
Georgia Institute of Technology

K. Warner Schaie
Pennsylvania State University

We analyzed data on psychometric intelligence from the Seattle Longitudinal Study, simultaneously estimating longitudinal factors, their covariance structure, and their mean levels. Data on five Thurstone Primary Mental Abilities subtests were available for 412 adults, ages 22–70 at first test, who were tested three times at 7-year intervals. A previous longitudinal factor analysis had shown high stability of individual differences (covariance stability) in general intelligence for three adult age groups. We extended that model to estimate factor means. All three age groups showed high levels of covariance stability, but differed sharply in their mean profiles. The young group showed increasing levels of general intelligence, the middle-aged group had stable levels of intelligence, and the old group showed salient, approximately linear, decline. The patterns of stability in middle-age, followed by mean decline and high covariance stability in old age, suggest a normative developmental transition from a stability pattern to a decline pattern of general intelligence, with the inflection point occuring somewhere around age 60.

An important issue in the study of adult intellectual development concerns whether levels of intelligence remain stable with advancing age. There is general agreement that the average level of performance on certain psychometric measures of intelligence declines with age, although there is great debate as to (a) the ubiquity of decline, (b) the proper interpretation of decline in psychometric performance, when it occurs, and (c) the practical importance of the magnitude of age-related decline (e.g., Baltes, Dittman-Kohli, & Dixon, 1984; Botwinick, 1977; Dixon, Kramer, & Baltes, 1985; Horn, 1985; Horn & Donaldson, 1976, 1980; Schaie, 1983). At the center of the disagreements in the literature regarding aging and intelligence has been Schaie's longitudinal studies of aging and primary mental abilities (see Schaie, 1983). The debate between Horn, Schaie, and others (e.g., Baltes & Schaie, 1976; Horn & Donaldson, 1976) covered a large number of issues associated with Schaie's sequential design, psychometric tests, and alternate theories and interpetations of aging and intelligence. Subsequent work by Schaie and Hertzog (1983) re-examined the issues with new data from Schaie's sequential samples. Their cohort-sequential analyses identified clear cohort differences in certain psychometric tests and identified statistically significant changes in multiple psychometrically defined abilities. For all five subtests of Thurstone's Primary Mental Abilities (PMA; Thurstone, & Thurstone, 1949), declines in performance (whether measured by longitudinal or cross-sectional sequences) were negligible until after age 50. Declines that were observed after age 50 were small, but became increasingly large after mean age 60. A somewhat surprising result, given earlier cross-sequential results from Schaie's data, was that the longitudinal sequences suggested decline after mean age 60 in all PMA subtests, although the decline began later for the PMA subtest Verbal Meaning (a test of recognition vocabulary). Schaie and Hertzog (1983) argued that these results required some minor modification of previous positions regarding the age of onset of intellectual decline, but that they supported the major conclusions of (a) age-confounded cohort differences in cross-sectional studies, (b) relative stability of mean performance levels into the 50s, with substantial declines *only* after age 60, and (c) some differences across subtests in the onset and magnitude of age-related performance declines (see also Dixon et al., 1985).

Although most of the gerontological literature has focused on the issue of stability of mean levels of intelligence with aging, *mean stability* is but one type of stability that can be assessed in longitudinal data. Another important type of stability is *stability of individual differences* (e.g., Baltes, Reese, & Nesselroade, 1977; Kagan, 1980; Schaie & Hertzog, 1985). This stability reflects the degree to which individuals differ in their developmental patterns of change (Baltes et al., 1977; Nesselroade & Labouvie, 1985; Schaie & Hertzog, 1985). Whereas stability of means is reflected in equivalent mean values at different developmental times, stability of individual differences is reflected in the covariance of a variable with itself over two points in time (see Baltes et al. 1977). In this article, we refer to stability of individual differences as *covariance stability* (see Hertzog & Nesselroade, 1987).

We report data collected as part of the Seattle Longitudinal Study, which has been supported over an extended period of time by grants from the National Institutes of Health, the National Institute of Child and Human Development, and the National Institute on Aging. Currently, the study is supported by Public Health Service Grant R01-AG04770 from the National Institute on Aging. The first author's effort on this project was also supported by Research Career Development Award K04-AG00335 from the National Institute on Aging.

The cooperation and support from members and staff of the Group Health Cooperative of Puget Sound is gratefully acknowledged.

Correspondence concerning this article should be addressed to Christopher Hertzog, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332-0170.

In a previous article, Hertzog and Schaie (1986) demonstrated that there is substantial covariance stability in intelligence across the adult life span. Hertzog and Schaie (1986) used a longitudinal factor analysis of data from the Seattle Longitudinal Study (SLS; Schaie, 1983) to show (a) that a general intelligence factor, *g*, could be identified for three age groups (young, middle-aged, and old), (b) that this *g* factor was defined equivalently by the PMA subtests in each age group and showed invariant factor loadings across longitudinal occasions, (c) that the covariance stability of *g* was high in all age groups, with longitudinal correlations of *g* with itself at or above .9 between successive longitudinal occasions, even in the older group, and (d) that there was substantial covariance stability in the five primary ability subtests, independent of *g*, as reflected in the proportion of variance in the PMA subtests determined by "test-specific" factors.

Hertzog and Schaie's (1986) results support the hypothesis that age changes in *g* are relatively consistent for same-aged individuals. Although there are individual differences in change patterns, these differences produce shifts in relative ordering of individuals that are small relative to the overall population variance in *g*. It is interesting that covariance stability was high in age ranges in which Schaie and Hertzog (1983) detected decline in the individual PMA subtests—namely, after age 60. This finding suggests only modest individual differences in the magnitudes of late-life decline in *g*.

We report a series of additional analyses designed to examine explicitly the mean level stability of *g* and, simultaneously, to estimate stability of individual differences in *g*. The results of these analyses demonstrate the independence of these two type of stability in the domain of psychometric intelligence. The analyses also were used to examine the question of inflection point for shifts from stability to decline in general intelligence.

The simultaneous examination of mean and covariance stability in longitudinal data is made possible by use of structural equation models to analyze means of latent variables (e.g., McArdle & McDonald, 1984; Sörbom, 1982). The longitudinal factor analyses reported by Hertzog and Schaie (1986) constitute an important precursor to simultaneous analysis of mean and covariance structures. Hertzog and Schaie found metric invariance in the *g* factor loadings between groups and across longitudinal occasions of measurement. Metric invariance is defined as equivalence in the unstandardized regression weights of variables on factors (see Horn, McArdle, & Mason, 1984). As discussed by several developmental methodologists (e.g., Baltes & Nesselroade, 1973; Labouvie, 1980a, 1980b; Schaie & Hertzog, 1985), an assumption of metric invariance is essential for allowing unambiguous interpretation of quantitative differences in mean levels of factor scores. The demonstration of metric invariance in *g* ensures that *g* is measured in equivalent units of measurement, so that differences in *g* factor means are uncontaminated reflections of mean level differences in the latent variable (see Labouvie, 1980a, 1980b; Schaie & Hertzog, 1985, for further discussion of this issue).

Given evidence of metric invariance, the simultaneous analysis of means and covariance structures requires introduction of the means into the structural equations of the longitudinal factor model already used by Hertzog and Schaie (1986). The critical questions of interest were (a) What is the magnitude of mean age changes in *g* at the different age levels studied? (b) Do age differences and age changes in *g* fully account for the mean changes in PMA subtests, or must different developmental trends of PMA means be modeled to account fully for the information in the means? and (c) Is there evidence for independence of stability of *g* means from the covariance stability of *g*?

Method

Subjects

The subjects in this study were participants in the Seattle Longitudinal Study conducted by Schaie and his associates (Schaie, 1983). The population consisted of members of a health maintenance organization (HMO) in the greater Seattle area. The population was defined as all of the members of the HMO as of 1956, the initial year of the longitudinal study, in order to minimize the probability of selection differences over time. All of the participants were unpaid volunteers who answered questionnaires and took part in a single psychometric test session. The participants, adults between the ages of 20 and 74 years at the first test, represented a range of socioeconomic and ethnic groups (although the population defined by the HMO membership in 1956 was predominantly White and somewhat more affluent than the general Seattle population). Further details on the population and sampling procedures may be found in Schaie (1983).

Sequential Sampling Design

The longitudinal samples studied here are a subset of the sequential samples collected in the SLS. The sampling plan of the SLS is discussed more fully in Schaie (1983), and the present sample is defined explicitly in Hertzog and Schaie (1986). Briefly, we restrict our analysis here to two 14-year longitudinal samples (first tested in 1956 or in 1963). Data from the two longitudinal sequences were partitioned into a hybrid sequential data matrix described in Table 1. The partitioned data matrix forms three age groups for simultaneous analysis.

Variables

As part of a larger psychometric battery, all of the subjects were administered the 1948 version of the SRA Primary Mental Abilities Test, Form AM 11-17 (Thurstone & Thurstone, 1949). The 1948 PMA includes five subtests, all of which are timed and have significant speed components in adult samples (see Schaie & Hertzog, 1983): (a) Verbal Meaning—a test of recognition vocabulary; (b) Space—a test of spatial relations requiring mental rotation of figures in a two-dimensional plane; (c) Reasoning—a test of inductive reasoning requiring recognition and extrapolation of patterns of letter sequences; (d) Number—a test of the ability to solve simple two-column addition problems quickly and accurately; and (e) Word Fluency—a test of the ability to retrieve words from semantic memory according to an arbitrary syntactic rule (words beginning with the letter *s*). Scoring followed the PMA manual: Verbal Meaning and Reasoning were scored in terms of the number of correct items, Space and Number were scored by subtracting incorrect items (comission errors) from the total number of correct items, and Word Fluency was scored by tallying the number of unique, admissable words generated during the allotted time.

Models and Statistical Procedures

The longitudinal factor model used is an application of a generic longitudinal model described in some detail by Joreskog and Sorbom (1977; see also Hertzog, in press; Horn & McArdle, 1980; Schaie & Hertzog, 1985). A detailed description of the model may be found in

Table 1
Reparameterized Sequential Sample for Multiple Group Analysis

Group/sample	Cohort (*M* birth year)	Age			*n*
		Occasion 1	Occasion 2	Occasion 3	
Group 1					
1	1931	25	32	39	21
1	1924	32	39	46	26
2	1938	25	32	39	22
2	1931	32	39	46	40
M		30	37	44	
Total					109
Group 2					
1	1917	39	46	53	27
1	1910	46	53	60	32
2	1924	39	46	53	51
2	1917	46	53	60	50
M		42	49	56	
Total					160
Group 3					
1	1903	53	60	67	28
1	1896	60	67	74	15
1	1889	67	74	81	13
2	1910	53	60	67	48
2	1903	60	67	74	18
2	1896	67	74	81	21
M		58	65	72	
Total					143

Hertzog and Schaie (1986). The model specified an occasion-specific *g* factor at each longitudinal occasion. The factor covariance matrix modeled the variances and covariances of *g* at the different occasions of measurement, and the residuals in the PMA subtests were modeled as having test-specific covariances (e.g., the residuals for Verbal Meaning were allowed to covary across longitudinal occasions). The specification of longitudinal models including factor means is relatively complex (Joreskog & Sorbom, 1984; McArdle & Epstein, 1987; Sorbom, 1982). The critical features are (a) a vector of location constants, analogous to grand means, (b) representation of latent variable means as regressions on a fixed constant and modeled in the LISREL GAMMA parameter matrix, and (c) the assumption that the means of all residuals are zero in the population. The vector of location constants identifies an intercept for each observed variable (PMA subtest). In longitudinal analysis of multiple groups, these location parameters are constrained equal both across longitudinal occasions and between the multiple age groups. Given data containing neither group differences nor longitudinal changes in means, this location parameter vector would perfectly account for the mean structure. Thus, the model with factor means will be meaningful only if there are either group differences or longitudinal changes in observed variable means that the model may attempt to structure as a function of the factor means.

Identification of the location parameters and the factor means is achieved by fixing the mean of *g* to zero for one age group at one longitudinal occasion. In the models reported, we fixed the *g* mean for the middle-aged group at the first occasion (mean age 42) at zero. This procedure then enables the remaining factor means to be estimated as deviations from this reference point (see Joreskog & Sorbom, 1984; Sorbom, 1982) for additional details. The fact that factor means are modeled as regression of factors (i.e., *g*) on a constant requires the assumption that the means of the residuals are zero. This is an unlikely assumption, given that we expect age trends in mean levels to vary across PMA subtests (independent of their relation to *g*). It is, however, possible to estimate residual component means by moving these parameters into the latent variable vector in LISREL.[1]

All of the models were estimated in either LISREL V or VI (Joreskog & Sorbom, 1984) using maximum likelihood estimation. In structural modeling, model fit can be assessed by likelihood ratio chi-square, as well as relative fit indices provided by the program. These indices are of less value in models with means, however, so we report a decomposition of overall model fit into (a) fit of the covariance structure model and (b) fit of the mean structure model (see Bentler & Bonett, 1980; Sobel & Bohrnstedt, 1985). The relative fit index for the means may be interpreted as an index of the proportion of information in the mean structure, adjusted for location parameters, accounted for by the model.

The procedures used here are unabashedly exploratory in nature. The goal is to use the LISREL model to explore descriptive developmental hypotheses about the longitudinal mean and covariance structures of the PMA subtests. This use of a generic longitudinal factor model is an appropriate application of structural equation techniques, which are ideal for exploratory multivariate modeling of longitudinal data (Hertzog, in press; McArdle & Epstein, 1987). This study cannot and should not be considered to represent a confirmatory analysis, in the philosophical sense of the term.

Results

The first model we estimated fixed the *g* factor means at zero in all three age groups, but allowed all location parameters to be freely estimated. This model fits the 15 means of each age group with 15 freely estimated location parameters. There is

[1] A listing of the LISREL VI specifications for models with factor and residual means is available from the first author.

Table 2
Goodness-of-Fit for Longitudinal Factor Model With Means

Model	χ^2	*df*	*F*[a]	*p*
M_s (saturated)	287.68	248	.352	.048
M_n (null in means)	642.02	288	.785	.000
M_1 (*g* factor means)	467.59	280	.572	.000
M_2 (*g* factor means, all 0 in middle-aged)	470.88	282	.575	.000
M_3 (*g* and test-specific factor means)	338.76	270	.414	.003
M_4 (*g* and residual means for V, S, N, W)	299.05	254	.366	.027

Note. V = Verbal Meaning, S = Space, R = Reasoning, N = Number, W = Word Fluency.
[a] LISREL fitting function at minimum.

a one-to-one correspondence between location parameters and sample means, and as such, the location parameters are just-identified. This model is therefore saturated with respect to the means, using Bentler and Bonett's (1980) definition. The fit of the model, denoted M_s, is reported in Tables 2 and 3. As expected, this model fit the same as the model ignoring means reported by Hertzog and Schaie (1986), and yielded an identical longitudinal factor solution. A second preliminary model, following recommendations of Bentler and Bonett (1980), was a null model in the means. This model specified five location parameters, one for each PMA subtest, and constrained these parameters to fit the means of all three longitudinal occasions for all three age groups. Thus, the 45 population means were fit with five location parameters. This null model, M_n, would have a fit equal to the saturated model, M_S, if there were no group differences or longitudinal changes in PMA subtest means to structure as part of the analysis. There was, however, a substantial, statistically significant difference between the two models, as seen in the first model comparison reported in Table 3. Clearly, there was longitudinal and age group variation in the PMA means, and the task of the analysis was to structure this variation in terms of the longitudinal factor model.

The first substantive model of interest specified *g* factor means in all three age groups. Interpretation of the fit of these substantive models must be made on the basis of relative differences from the null and saturated models, so that one can evaluate fit to the means ignoring (assuming) the basis specification and fit of the longitudinal factor model (Bentler & Bonett, 1980; Sobel & Bohrnstedt, 1985). In essence, the difference between the null and saturated models defines a range of possible fits of models structuring means in the longitudinal analysis. The critical question is how close a model with structured means comes to the fit of the model that is saturated in the means (or conversely, how far it has come from the poor fit of the null model).

As shown in Table 3, this first substantive model, M_1, improved meaningfully on the fit of the null model, although there was still a significant difference between M_1 and M_S. The relative fit of the new model is best indexed by the Sobel and Bohrnstedt (1985) relative fit index, denoted as δ in Table 3. The fit of .49 indicates that about half of the variation in the means had successfully been structured by M_1.

One interesting outcome of model M_1 was that the *g* factor means for the middle-aged adults were not significantly different from zero, relative to their standard errors. In models of this type, these estimated factor means are scaled as deviations from the fixed zero mean (age 42 for the middle-aged population). Therefore, the finding of essentially zero *g* means at ages 49 and 56 for the middle-aged group indicated no statistically significant change in mean level of *g* over this age range. A second model, M_2, incorporated this feature by fixing the *g* means to zero for all three ages of the middle-aged group. This model did not fit more poorly than M_1.

The fact that M_2 fit significantly worse than M_S implied that the assumption of no mean variation in the residuals for the PMA factors had to be abandoned. That is, it was not possible to model age-group differences and age changes in PMA means solely as a function of age differences and age changes in *g* factor means. Apparently, the primary abilities measured by the PMA have variations in the means that are saliently different from the behavior of the *g* factor means.

A logical possibility is that there are age group differences in subtest-specific means, but no age group differences in patterns

Table 3
Comparisons of Fit Between Alternative Models With Factor Means

	M_s		M_n				Comparison		
Model	$\Delta\chi^2$[a]	Δdf	$\Delta\chi^2$[b]	Δdf	δ[c]	Comparison	$\Delta\chi^2$	Δdf	$\Delta\delta$[d]
M_s	—	—	—	—	—	—	—	—	—
M_n	—	—	—	—	—	M_n–M_s	354.34	40	—
M_1	174.43	8	179.91	32	.492	—	—	—	—
M_2	171.94	6	182.40	34	.485	M_1–M_2	2.49	4	.007
M_3	303.24	18	51.08	22	.857	M_2–M_3	128.83	10	.365
M_4	342.97	34	11.37	6	.968	M_1–M_4	168.54	28	.483

[a] Difference in χ^2 between model and M_n (null model).
[b] Difference in χ^2 between model and M_s (saturated model).
[c] Relative fit index for fit to the mean structure.
[d] Change in relative fit index in means for models under comparison.

of age changes in the primary ability means. Such a pattern could arise if age changes in the primary abilities were solely a function of age changes in *g*, but there were also differential patterns of cohort effects across the primary ability means. Our previous work (Hertzog & Schaie, 1986), modeling both *g* and PMA test-specific factors, provided a convenient means of testing this hypothesis. We used a model that specified eight factors in each age group: (a) three *g* factors, one at each longitudinal occasion, and (b) five test-specific factors, one for each PMA subtest. We estimated factor means for all eight factors, achieving identification of the test-specific factor means by fixing all five test-specific factor means for the middle-aged group to zero. This model, M_3, allowed the *g* factor means at ages 49 and 56 to be freely estimated in the middle-aged group, as in model M_1. We did not wish to assume mean stability in *g*, even though that was suggested from the M_2–M_1 comparison. It could have been the case that the stable *g* factor means in the middle-aged group in the previous models were an artifact of model misspecification.

Model M_3 also constrained the test-specific factor loadings to be equal over the three age groups (see Hertzog & Schaie, 1986). The equality constraints on test-specific factor loadings did not permit any of the age-group differences in mean changes to be modeled by the test-specific factor means. Group differences in mean change on the PMA variables could only be reflected in the *g* factor means.

Table 2 reports the fit of M_3. The model fit significantly better than M_1, indicating there were statistically significant age group differences in test-specific factor means. However, the model still did not approximate the fit of M_5, requiring rejection of Model M_3. It was also still the case that the *g* factor means did not differ significantly between ages 42 and 56 for the middle-aged group. We concluded that there were age-group differences in PMA subtest means, but that there are also differential age changes for the PMA subtest means, independent of *g*. We also concluded that it was still plausible to maintain the assumption of no age changes in *g* in the middle-aged group.

We next proceeded by fitting a series of models allowing residual means. This approach was needed to allow for age-group differences in patterns of mean age changes on the primary abilities. This series of models proceeded in exploratory fashion. Large mean residuals (differences between sample means for the PMA subtests and PMA means predicted from the model parameters) and salient LISREL modification indices were used to indicate a need for structuring additional mean parameters. Unlike M_3, these models specified a separate PMA residual "factor" at each longitudinal occasion, permitting both *g* and the PMA residuals from *g* to display age-related change. After a series of model modifications, we arrived at a model that did not differ significantly from the saturated model. This model allowed residual means for Word Fluency, Number, Verbal Meaning, and Space. This modified model, M_4 in Table 2, achieved a relative fit index of .97 to the means, indicating excellent fit. Of course, this fit was achieved by adjusting to the sample means, and can therefore be treated only as a descriptive index of the success of the model modification process.

One of the major reasons for fitting additional models to the means was to ensure that the estimated age changes and age differences in *g* means were not inappropriately biased by the incorrect assumption of no residual means. Hertzog and Carter (1982) previously demonstrated that group differences in intelligence factor means were affected by the specification error of zero residual means. Table 4 reports the *g* factor means for the four substantive models, M_1 through M_4. Irrespective of the model, the relative pattern of *g* factor means in the three age groups remained the same. The *g* factor means increased from mean age 30 to mean age 37 in the young group, and then remained relatively stable through age 44. The *g* factor exhibited mean stability from mean age 42 through mean age 56 in the middle-aged group. Finally, *g* showed substantial decline from mean age 58 through mean age 72 in the old group. The mean decline in *g* in the old group was roughly linear over the 14-year period. The comparable pattern of *g* mean behavior is particularly important in Model M_4, in which it was most likely that the apparent age changes in *g* estimated in Models M_1 through M_3 would change as a function of specifying longitudinal changes in the PMA residuals as well. The fact that conclusions regarding the behavior of *g* means were not altered by specifying longitudinal variation in PMA residual means indicated that the mean patterns were unlikely to be an artifact of model specification.

Approximate 99% confidence intervals around the factor means can be calculated by subtracting and adding 2.5 *SE*s to the estimated *g* factor means. Inspection of Table 4 clearly showed that these 99% confidence intervals did not include zero for any of the freely estimated means in the old and young groups. As these means are deviation contrasts from the middle-aged *g* means, we concluded there were reliable age group differences in means. The significant differences included comparisons between the different groups at roughly comparable ages. That is, the young group at age 44 (Occasion 3) differed significantly from the middle-aged group at age 42 (Occasion 1), as did the middle-aged group at mean age 56 (Occasion 3) from the old group at mean age 58 (Occasion 1). Although the hybrid sequential design does not completely unconfound age changes and cohort differences, it seems likely that these differences reflect cohort differences in the mean levels of *g*.

Table 5 reports the residual means estimated in the final model, M_4. These means must be interpreted with care. They represent mean patterns in the PMA subtests orthogonal to the trends mediated through *g*. The first feature of note involves the residual means for Word Fluency and Number in the middle-aged group. Although the *g* means showed no age-related changes in the middle-aged, the residuals for Word Fluency and Number did change. There were small but statistically significant declines in Word Fluency and Number between mean ages 42 and 56. There is a second noteworthy feature of the residual means in Table 4. It seems that the large age-group (cohort) differences in *g overestimated* age group differences in Number and Verbal Meaning. This was shown by the large negative means in the young group for these two PMA subtests, as well as the large positive means for Number for the old group. Finally, there appeared to be modest levels of decline in Space for the old group (between mean ages 58 and 65) that was greater than the decline in Space predicted by *g*.

We do not report here the other parameter estimates from the longitudinal solution (e.g., factor covariances, factor loadings) because they differed trivially from the solution without means

Table 4
The g Factor Means for Alternative Longitudinal Models

Group	*M* age	Model							
		M_1		M_2		M_3		M_4	
		M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Young									
g_1	30	1.61	0.60	1.62	0.59	8.54	3.26	2.82	0.65
g_2	37	2.76	0.57	2.78	0.57	10.11	3.49	3.99	0.65
g_3	44	2.70	0.56	2.71	0.55	9.87	3.39	3.50	0.62
Middle-aged									
g_1	42	0*	—	0*	—	0*	—	0*	—
g_2	49	0.10	0.17	0*	—	0.14	0.16	0*	—
g_3	56	−0.20	0.18	0*	—	−0.20	0.17	0*	—
Old									
g_1	58	−3.96	0.61	−3.97	0.60	−10.96	4.48	−4.20	0.64
g_2	65	−4.61	0.61	−4.62	0.61	−12.41	4.64	−4.78	0.64
g_3	72	−6.55	0.65	−6.57	0.64	−13.28	4.24	−6.22	0.66

Note. Asterisks denote fixed factor means. The *g* factor subscripts denote longitudinal occasion.

reported by Hertzog and Schaie (1986). However, one question remained regarding the factor covariance matrix for *g*. As reported in Hertzog and Schaie, there was an age-related increase in *g* factor variance in the old group. The old group also had greater overall variance in *g* than did the middle-aged and young groups. One possible explanation of these differences is that they are methodological artifacts. The old group was formed by pooling over a larger age span in order to achieve acceptable sample size for structural analysis (refer back to Table 1). In the present context, it was possible that the developmental changes in *g* factor means would differ if the youngest age group (mean age 53 at Occasion 1; age range, 50 to 56) were omitted from the analysis. To address this question, we redefined the old group to include only the individuals age 57 and older at first test, and re-ran the longitudinal model with this subsample. Briefly, this analysis showed (a) similar age declines in *g* means, but of greater magnitude, (b) higher variability in *g* in the old group, but (c) more homogeneity of *g* variance across the three longitudinal occasions. Thus, it appears that the increasing variability in *g* over time, found in the full sample, reflected differences in developmental patterns from ages 50 to 65, as opposed to heterogeneity of developmental trajectories of same-aged individuals in the latter part of the adult life span. The analysis thus provides further support for the argument of an inflection point around age 60, at which age decrements in PMA performance begin to accelerate. The increased variability in *g* in the older group is not, however, merely a methodological artifact of age-group definition.

Table 5
Residual Means in Final Model (M_4)

Variable Occasion	Age group					
	Young		Middle-aged		Old	
	M	*SE*	*M*	*SE*	*M*	*SE*
Verbal Meaning						
1	−5.30	1.01	0*		0.26	0.98
2	−4.23	1.07	0*		1.09	1.05
3	−3.65	1.03	0*		−0.49	1.08
Space						
1	0.58	1.15	0*		−1.19	1.01
2	0.98	1.22	0*		−2.68	1.01
3	1.76	1.20	0*		−2.56	1.03
Reasoning						
1	0*		0*		0*	
2	0*		0*		0*	
3	0*		0*		0*	
Number						
1	−5.56	1.32	0*		3.71	1.23
2	−5.58	1.40	0.28	0.44	5.12	1.28
3	−6.03	1.31	−1.62	0.43	3.38	1.27
Word Fluency						
1	−1.45	1.48	0*		4.98	1.45
2	−3.56	1.59	−1.43	0.68	2.77	1.46
3	−1.18	1.60	−2.08	0.69	2.36	1.49

Note. Asterisks denote fixed 0 parameters.

Discussion

The results from this analysis amplify and accentuate several issues regarding age changes in psychometric intelligence. First, the results extend Schaie's (1983) work on age patterns in multiple primary intellectual abilities to the level of general intelligence, as measured by the *g* factor defined from the PMA subtests. We found a pattern of age changes in *g* factor means highly consistent with previous univariate results (e.g., Schaie & Hertzog, 1983). There were small increases in *g* in early adulthood (through mean age 32), stability in *g* means through middle age (until mean age 56), and substantial decline in late life. We explicitly tested the hypothesis that there was no decline in *g* in the middle-aged group at two different junctures, and could not reject the hypothesis. Moreover, the age changes that were estimated as part of this hypothesis test were so small as to be trivial in importance. On the other hand, we did find evidence of some

decline in the middle-aged group on the PMA subtests Word Fluency and Number, independent of *g*.

The results also suggest substantial cohort differences in *g* means. The age groups differed not only in terms of mean age at initial test but also in birth cohort membership. The fact that the middle-aged group at mean age 56 performed significantly better on *g* than did the old group at mean age 58 surely indicates salient cohort differences in these data, as already detailed by Schaie (1983).

The unique contribution of this study, in terms of estimating age changes in PMA means, stems from the fact that the mean differences are estimated at the level of the *g* factor. Because these estimates are based on the simultaneously estimated factor pattern weights, they represent optimal estimates of *g* factor means that are not contaminated by mean patterns specific to the primary abilities themselves. Moreover, the analysis permitted the evaluation of mean trends in the primary abilities after they have been residualized with respect to *g*.

An additional contribution of the present analysis is that it permits independent evaluation of mean stability and covariance stability in *g*. These results demonstrate concretely the independence of these two types of stability. In all three age groups, individual differences in *g* were highly stable over the 14-year period. Yet each age group showed dramatically different age trends in *g*. In the young group, *g* increased to a stable plateau. In the middle-aged group, *g* means remained stable, but in the old group, substantial *g* decline was observed.

The change in mean patterns across the age groups, coupled with the high degree of covariance stability across the life span, has important implications for several prominent hypotheses about adult intellectual development. It is often the case, especially recently, that *g* is identified with basic intelligence (e.g., Jensen, 1982). Given (a) the widely accepted notion that there is multidirectionality in age trends in ability, such that some, but not all, abilities show age-related declines (e.g., Baltes et al., 1984; Botwinick, 1977; Horn & Donaldson, 1980) and (b) the accepted argument that it is measures of fluid intelligence (Horn, 1985; Horn & Donaldson, 1980), or alternatively, Wechsler-type performance tests (Botwinick, 1977; Salthouse, 1982) that manifest early decline, one would expect that *g*, as measured here, would be the prime candidate for evidencing decline from ages 25 to 55. To the contrary, it appears to be the case that *g* manifests *both* mean stability and covariance stability in middle age in the Seattle Longitudinal Sample.

How can this discrepancy be explained? One possible explanation is that the *g* factor estimated by the PMA variables is highly specific to the variables or to the samples, and hence is in some way a poor measure of the construct of general intelligence. This possibility seems relatively implausible. The *g* factor loadings estimated here are highly consistent with those found by Thurstone and Thurstone (1941) for these tests, and show a pattern of loadings consistent with a plethora of studies from the psychometric literature. The best indicator of *g* in the PMA, judged from our factor loadings, is Reasoning. This subtest, a measure of induction, is probably the best indicator of general intelligence and of the Horn-Cattell second-order fluid intelligence factor in the PMA (Horn & Donaldson, 1976). Not only did the Reasoning test load highly on *g*, but the Reasoning means in all age groups were well fit by the models specifying no age-related changes in *g* in the middle-aged group. Although we have estimated the single higher order *g* factor here, as opposed to fluid intelligence, Gustaffson (1984) recently reported hierarchical factor results from multiple intelligence tests that suggest that the *g* factor is isomorphic with fluid intelligence.

Thus, it would seem that the hypothesis of early decline in *g* is not supported by these data. The best model for the development of *g* in middle-age is a model of stability in both means and individual differences. One could argue that the generalizability of these results is limited because individuals who manifest early decline are more likely to drop out of longitudinal studies. Perhaps so, but the finding of mean stability of *g*, even in a select subpopulation, argues against the ubiquity of early age declines in *g*. There is evidence in these data of decline in two PMA subtests, Word Fluency and Number, in the middle-aged group. We suggest that, barring the sort of nonnormative events that lead to early mortality, individuals appear to maintain stable performance levels of *g* until sometime after age 50.

However, the developmental pattern of *g* begins to change dramatically between ages 50 and 60. After mean age 58, we found substantial, statistically significant decrements in mean levels of *g*. This decline was observed in an age group in which the covariance stability of *g* remained quite high. These results, then, offer little support to the hope that age-related decline in *g* is somehow nonnormative or is restricted to a small subpopulation of older individuals. We did find increased variance in *g* in the middle-aged and older groups, suggesting some small differences in developmental trajectories between those individuals in their 50s and those in their 60s. However, the longitudinal increases in *g* variance in the older group—crucial to the argument of different developmental trajectories in old age—were eliminated when the old group was restricted to individuals age 57 and older at first test.

The fact that it was necessary to fit residual mean factors, varying in age patterns, provides support for the arguments of Baltes and colleagues (e.g., Baltes et al. 1984) that intelligence is both multidimensional and multidirectional in its development. For example, the fact that young adults have *lower* means on the Verbal Meaning residuals suggests that the *g* factor means overestimate the age differences in vocabulary, even though Verbal Meaning has high loadings on *g*. This pattern is also observed for the Number and Word Fluency residual means, and may suggest reversed cohort differences on these tests when *g* is statistically removed from these tests. The pattern of Space residual means in the old group indicates greater decline between ages 58 and 65 on spatial ability than is predicted by *g*. Some caution is in order in interpreting these residual means. Our data only permit estimation of factor means for *g*. These residual means do not have the same status as means estimated in models with multiple measures of each primary ability, being much more likely to be specific to the PMA subtest than would primary ability factor means.

The analysis provides relatively little evidence of substantial individual differences in intraindividual change in general intelligence. To the contrary, these findings of differential age group patterns in *g* means, coupled with high degree of covariance stability in all age groups, suggest a relatively *normative* developmental transition in *g*. That is, it appears that most individuals make a transition from a stability to a decline pattern of *g*

development at some point between age 55 and age 70, with individual differences in the age of onset of this transition.

It is important to note that these inferences are based on population parameters, and that there are some individuals who do not show salient decline even into old age (Schaie, 1983). There may be greater heterogeneity of change for the primary abilities, as opposed to *g* (see Hertzog & Schaie, 1986). Nevertheless, the results suggest that the heterogeneity of developmental trends in *g* during old age is small when measured against the population variance.

The high degree of covariance stability is a descriptive phenomenon and should not be assumed to demonstrate the validity of biological causes of age changes in *g*. Stability does not imply immutability, and Schaie and Willis (1986) have demonstrated significant training gains in inductive reasoning in individuals with prior histories of decline in this ability (all of whom were, in fact, part of the samples used in the present analysis).

In a sense, these results contradict aspects of the arguments made by both sides of the debate regarding the nature of intellectual decline manifested in the Seattle Longitudinal Study (Baltes & Schaie, 1976; Horn & Donaldson, 1976). The results appear, however, consistent with the updated perspectives of both Horn (1985) and Baltes and his colleagues (e.g., Baltes et al., 1984; Dixon et al., 1985). The key involves an assessment of the kinds of abilities measured in timed psychometric tests such as the Thurstone PMA, and hence, the nature of the *g* factor extracted from it. Evidence from a number of studies have shown that Thurstone-type tests of primary abilities have high correlations with speed of basic perceptual processes in adult samples (Cornelius, Willis, Nesselroade, & Baltes, 1983; Hertzog, 1987; Horn, Donaldson, & Engstrom, 1981). Schaie originally selected the adolescent form of the PMA for his study, and this form has limited item difficulty and substantial speed components in adult samples (e.g., Schaie, Rosenthal, & Perlman, 1953). The *g* factor estimated in this study was marked as highly by PMA Verbal Meaning as by PMA Reasoning. We have recently shown a strong relationship of PMA Verbal Meaning to a Perceptual Speed factor independent of its relationship to other vocabulary tests (e.g., ETS Advanced Vocabulary; Schaie, Willis, Hertzog, & Schulenberg, 1987). Thus, it appears that the PMA was constructed so as to maximize variance determined by what might be termed the *mechanics* of intelligence (e.g., Hunt, 1978), that is, the speed of basic cognitive processes needed for rapid decisions of low to moderate difficulty. Given that age-related slowing in information-processing speed is a highly normative developmental phenomenon (e.g., Birren, 1974; Salthouse, 1985), we can construct the following argument. The PMA manifests little age change in *g* prior to age 55 because *g* as operationally defined by the PMA, emphasizes speeded solution of problems of limited difficulty. However, sometime after age 50, the age-related slowing in information-processing speed becomes a salient limiting factor in PMA performance, and *g* begins to decline dramatically. Individual differences in decline are minimized because (a) the PMA items are not optimally sensitive to the type of cognitive processes likely to maximize psychometric test performance in superior old adults (e.g., strategies for solving difficult problems, cognitive styles, and metacognitive processes; Baron, 1985; Dixon, in press; Sternberg, 1985) and (b) the ability domain covered by the tests is highly limited, excluding the types of abilities most likely to show increment and differential growth in adulthood, such as social cognition, domain-specific procedural knowledge, expertise, and postformal reasoning (Berg & Sternberg, 1985; Dixon et al., 1985; Labouvie-Vief, 1985; Rybash, Hoyer, & Roodin, 1986). Although important gains can be made by studying these other domains of cognition, we maintain that the study of cognitive mechanics, as they relate to performance on intelligence tests, remains a continuing priority for gerontology. A formal test of the cognitive mechanics interpretation of psychometric test performance in adulthood requires investigation of the nature of the information-processing skills tapped by Thurstone-type tests, research now ongoing in several laboratories.

References

Baltes, P. B., Dittmann-Kohli, F., & Dixon, R. A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 6, pp. 34–76). New York: Academic Press.

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 219–252). New York: Academic Press.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks-Cole.

Baltes, P. B., & Schaie, K. W. (1976). On the plasticity of intelligence in old age: Where Horn and Donaldson fail. *American Psychologist, 31*, 720–725.

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Berg, C. A., & Sternberg, R. J. (1985). A triarchic theory of intellectual development during childhood. *Developmental Review, 5*, 334–370.

Birren, J. E. (1974). Translations in gerontology—from lab to life: Psychophysiology and the speed of response. *American Psychologist, 29*, 808–815.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren & K. W. Schaie, (Eds.), *Handbook of the psychology of aging* (pp. 580–605). New York: Van Nostrand Reinhold.

Cornelius, S. W., Willis, S. L., Nesselroade, J. R., & Baltes, P. B. (1983). Convergence between attention variables and factors of psychometric intelligence in older adults. *Intelligence, 7*, 253–269.

Dixon, R. A. (in press). Questionnaire research on metamemory and aging: Issues of structure and function. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adulthood and old age*. New York: Cambridge University Press.

Dixon, R. A., Kramer, D. A., & Baltes, P. B. (1985). Intelligence: Its life-span development. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurement, and applications* (pp. 469–518). New York: Wiley.

Gustafsson, J. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179–203.

Hertzog, C. (1987). *The influence of cognitive slowing on age differences in intelligence*. Unpublished manuscript.

Hertzog, C. (in press). On the utility of structural regression models for developmental research. In P. B. Baltes, D. Featherman, & R. M. Lerner (Eds.), *Life-span development and behavior* (Vol. 10). Hillsdale, NJ: Erlbaum.

Hertzog, C., & Carter, L. (1982). Sex differences in the structure of intelligence: A confirmatory factor analysis. *Intelligence, 6*, 287–303.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait–state distinction for the structural modeling of developmental change. *Child Development, 58*, 93–109.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. *Psychology and Aging, 1*, 159–171.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theory, measurements, and applications* (pp. 267–300). New York: Wiley.

Horn, J. L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. *American Psychologist, 31*, 701–719.

Horn, J. L., & Donaldson, G. (1980). Cognitive development in adulthood. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 445–529). Cambridge, MA: Harvard University Press.

Horn, J. L., Donaldson, G., & Engstrom, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. *Research on Aging, 3*, 33–84.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. W. Poon (Ed.), *Aging in the 1980's: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.

Horn, J. L., McArdle, J. J., & Mason, R. (1984). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1*, 179–188.

Hunt, E. (1978). The mechanics of verbal ability. *Psychological Review, 85*, 109–130.

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). New York: Springer.

Joreskog, K. G., & Sorbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D. S. Aigner & A. S. Goldberger (Eds.), *Latent variables in socio-economic models* (pp. 285–325). Amsterdam: North Holland.

Joreskog, K. G., & Sorbom, D. (1984). *LISREL VI user's guide*. Mooresville, IN: Scientific Software.

Kagan, J. (1980). Perspectives on continuity. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 26–74). Cambridge, MA: Harvard University Press.

Labouvie, E. W. (1980a). Identity versus equivalence of psychological measures and constructs. In L. W. Poon (Ed.), *Aging in the 1980's: Psychological issues* (pp. 493–502). Washington, DC: American Psychological Association.

Labouvie, E. W. (1980b). Measurement of individual differences in intraindividual changes. *Psychological Bulletin, 88*, 54–59.

Labouvie-Vief, G. (1985). Intelligence and cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 500–530). New York: Van Nostrand Reinhold.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110–133.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37*, 234–251.

Nesselroade, J. R., & Labouvie, E. W. (1985). Experimental design in research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 35–60). New York: Van Nostrand Reinhold.

Rybash, J. M., Hoyer, W. J., & Roodin, P. A. (1986). *Adult cognition and aging: Developmental changes in processing, knowing, and thinking*. New York: Pergamon Press.

Salthouse, T. A. (1982). *Adult cognition: An experimental psychology of human aging*. New York: Springer.

Salthouse, T. A. (1985). *A theory of cognitive aging*. Amsterdam: North Holland.

Schaie, K. W. (1983). The Seattle Longitudinal Study: A 21-year exploration of psychometric intelligence in adulthood. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 64–135). New York: Guilford Press.

Schaie, K. W., & Hertzog, C. (1983). Fourteen-year cohort-sequential analyses of adult intellectual development. *Developmental Psychology, 19*, 531–543.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 61–92). New York: Van Nostrand Reinhold.

Schaie, K. W., Rosenthal, F., & Perlman, R. M. (1953). Differential mental deterioration on factorially "pure" functions in later maturity. *Journal of Gerontology, 8*, 191–196.

Schaie, K. W., & Willis, S. L. (1986). Can decline in intellectual functioning be reversed? *Developmental Psychology, 22*, 223–232.

Schaie, K. W., Willis, S. L., Hertzog, C., & Schulenberg, J. E. (1987). Effects of cognitive training on primary mental ability structure. *Psychology and Aging, 2*, 233–242.

Sorbom, D. (1982). Structural equation models with structured means. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (pp. 183–195). North Holland: Amsterdam.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). San Francisco, CA: Jossey-Bass.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of intelligence*. New York: Cambridge University Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence* (Psychometric Monographs, No. 2). Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1949). *Examiners manual: SRA Primary Mental Abilities Test* (Form 11–17). Chicago: Science Research Associates.

Received January 12, 1987
Revision received July 15, 1987
Accepted July 17, 1987 ■

APPENDIX F

TEXT RECALL IN ADULTHOOD: THE ROLE OF INTELLECTUAL ABILITIES

By

David F. Hultsch
Christopher Hertzog
Pennsylvania State University

Roger A. Dixon
Berlin, West Germany

Developmental Psychology
1984 Vol. 20, No. 6, 1193-1209

Text Recall in Adulthood: The Role of Intellectual Abilities

David F. Hultsch and Christopher Hertzog
Department of Individual and Family Studies, Pennsylvania State University

Roger A. Dixon
Max Planck Institute for Human Development and Education, Berlin, West Germany

This study examined age-related predictive relationships between an array of psychometric intellectual ability markers and text recall performance in adulthood. One hundred and fifty women from three age groups (21–39 years, 40–58 years, 60–78 years) read and recalled four narrative stories at three delay intervals and completed a battery of 12 factor-analytically defined intellectual ability tests. The results indicated (a) that text memory performance in adulthood is predicted by multiple abilities; (b) that age differences in text memory performance overlap highly with age differences in multiple abilities, although the latter do not fully account for the former; (c) that modest Age × Ability interactions exist but are not consistent with previous reports, suggesting that age differences decrease with increasing ability levels; and (d) that the pattern of intelligence-text recall relationships differs by age group.

Research examining the development of adult memory has shown that the existence of age-related differences in secondary memory performance is widespread (Craik, 1977; Poon, Fozard, Cermak, Arenberg, & Thompson, 1980). With few exceptions, younger adults routinely outperform older adults when the focus of the task is on verbatim recall of lists of numbers, symbols, words, and so forth. However, when the focus of the task is on the gist recall of meaningful, presumably ecologically valid text materials, the nature and extent of age-related performance differences are considerably less clear. A number of recent studies have reported age-related deficits in text processing that conform to the general pattern observed in verbatim recall of word lists (Cohen, 1979; Dixon, Simon, Nowak, & Hultsch, 1982; Taub, 1975, 1976; Taub & Kline, 1978; Zelinski, Gilewski, & Thompson, 1980). Other recent studies have found that younger and older adults appear to be equally adept at comprehending and remembering texts (Harker, Hartley, & Walsh, 1982; Meyer & Rice, 1981).

More specifically, the presence or absence of adult age differences in text processing appears to depend on multiple contextual factors (see reviews by Hultsch & Dixon, 1984; Meyer & Rice, 1983) including those related to the task (e.g., recall, recognition), materials (e.g., physical structure, organizational structure), and subjects (e.g., abilities, interests). For example, Simon, Dixon, Nowak, and Hultsch (1982) found middle-aged and older adults to be disadvantaged relative to young adults when asked for incidental recall of a text following performance of deep orienting tasks, however, they performed equally well following performance of a shallow orienting task and under intentional recall conditions. Similarly, in the case of material variables, Dixon, Hultsch, Simon, and von Eye (in press) found that age-related differences in the discovery and use of the organizational structure of texts depend, in part, on the number of concepts introduced in the text.

Although task and material variables play an important role in accounting for adult age-related performance differences in text processing, a major portion of the variance may be mediated by subject variables. For instance, it is reasonable to expect that indi-

This research was supported by research grant 1R01 AG00910-02 from the National Institute on Aging to David F. Hultsch and by predoctoral fellowship T 32 AG00049 to Roger A. Dixon.

Requests for reprints should be sent to David F. Hultsch, who is now at the Department of Psychology, University of Victoria, Victoria, British Columbia V8W2Y2, Canada.

vidual differences in education and verbal ability predict performance differences in text recall and that such individual difference variables may be related to the presence or absence of age differences in text recall performance (Meyer & Rice, 1983; Taub, 1979). In this context, it may be noted that studies reporting age differences in text recall have generally tested subjects with relatively low levels of education and verbal ability (i.e., high school graduates), whereas studies reporting little or no age differences have generally tested subjects with relatively high levels of education and verbal ability (i.e., college graduates).

In a recent analysis, Meyer and Rice (1983) examined text recall for four types of education/verbal ability subsamples drawn from a large sample of over 300 adults who had read and recalled two texts. Their analyses clearly indicated age differences favoring younger adults in populations with below average or average verbal ability and little education beyond high school. However, they did not find unequivocal evidence of age differences in subjects with higher levels of verbal ability and education. On the one hand, comparison of randomly selected younger adults and high-verbal older adults showed no significant age-related differences in performance. On the other hand, comparison of high-verbal younger adults and high-verbal older adults revealed age-related performance differences in favor of the young.

Similarly, Dixon et al. (in press) found that verbal ability appears to mediate age-related differences in the discovery and use of the organization of texts. Specifically, in the case of adults with relatively low levels of verbal ability, age-related differences in recall were greatest for the main ideas of the text. Younger and middle-aged groups did not differ significantly in recall of the details of the texts. However, both younger groups recalled significantly more details than the older adults. Thus, low-verbal older adults showed a deficit in recall of both the main ideas and the details of the text, although the size of the deficit was greater at the level of main ideas than at the level of details. In contrast, in the case of adults with high levels of verbal ability, age differences in recall were greatest for the details of the texts. There were no significant differences among the three age groups in the recall of the main ideas.

Although results like those of Meyer and Rice (1983) and Dixon et al. (in press) suggest that age differences in text recall interact with level of verbal ability, there are limitations to inferences drawn from extreme groups designs in which subjects are grouped according to extreme scores (e.g., upper vs. lower quartile) on a continuous variable. Age-related selection in the population makes it difficult to equate age/cohort groups partitioned on variables such as educational attainment and verbal ability (Krauss, 1980; Meyer & Rice, 1983). At a given level of education, a sample of older adults is probably more highly selected than a sample of younger adults because of cohort-related differences in educational attainment. Similarly, at a given level of verbal ability, a sample of older adults is probably less highly selected than a sample of younger adults because of age-related changes in vocabulary. In any case, it is virtually impossible to disentangle selection confounds from age differences produced by the aging process, even though the latter source of variance is obviously the one of interest.

There are other potential problems with the extreme-groups approach. The extreme-groups design ignores strength of prediction in the inner quartiles of the variables distributions. One could conclude that an Age × Ability interaction in an extreme-groups design indicated progressively smaller age differences with increasing ability levels, when in fact the age differences were consistent at all but the very highest levels of ability. A potential overgeneralization of extreme-groups interactions with age can only be avoided by examining the interaction across the full range of the ability distribution. An additional problem is that group assignment to extreme groups on the basis of scores on a single fallible variable may cause measurement error to have an unacceptedly high influence on the group assignment. Finally, other intellectual abilities and individual differences variables may mediate age differences in text processing. A comparison of groups differing on a single ability, however well measured,

cannot address the determination of individual differences in text processing by a well-defined domain of abilities.

The present study was designed to examine relationships between text recall performance, age, and a selected set of psychometric intellectual abilities using a multivariate correlational approach. More specifically, we sought to relate text recall performance at three delay intervals to a set of primary mental ability factors of intelligence (selected on the basis of their potential theoretical relevance to text processing): Induction, Memory Span, Associative Memory, Associational Fluency, Ideational Fluency, and Verbal Comprehension.

The present analysis has two major foci. First, we wished to determine whether there is an interaction between multiple intellectual abilities and age in determining individual differences in text recall performance, thus extending the logic of the previous extreme-groups studies to a regression analysis of the interactive relationship. Our analysis addresses the potential deficiencies in the extreme-groups paradigm by (a) examining the interaction at the level of intellectual ability *factors* rather than single ability variables and (b) producing product variable interaction terms that examine the interaction of ability and age across the range of the continuous distributions of abilities rather than in extreme groups. The regression analysis thus allows us to determine whether age differences in text recall are statistically independent of age differences in intellectual abilities, while analyzing whether any statistically independent age differences are qualified by the existence of ability/age interactions. Based on previous studies, we predicted that there would be Age × Ability interactions in text memory performance, with smaller age differences at higher ability levels.

The second focus of the study involved an analysis of individual differences in text recall/intellectual ability relationships *within* each age group. We predicted that the patterns of text recall–intelligence correlations would vary with age and the length of the recall delay interval. If true, these predictions would indicate an important qualification to any interpretation of Age × Ability interactions in text performance, because different abilities might be important for performance at different stages of the life span.

Method

Subjects

The subjects were 150 community-dwelling white women from a small city in central Pennsylvania. They were recruited through the Altoona campus of The Pennsylvania State University and local organizations such as churches and senior citizen centers. The subjects were paid $15.00 for their participation in the study.

The sample was divided into three age groups of 50 individuals. The youngest group ranged in age from 21 to 39 years (M = 32.02), the middle group ranged in age from 40 to 58 years (M = 49.48), and the oldest group ranged in age from 60 to 78 years (M = 68.96). The three age groups differed significantly in years of education (young: M = 13.62; middle-aged: M = 12.78; old: M = 10.98), $F(2, 147) = 14.77$, $p < .001$. In order to examine these differences further, the sample was broken down into semi-decade age groups, and the median educational attainment of these age groups was compared to that reported for these cohorts by the U.S. Bureau of the Census (1977). These comparisons suggested that the subjects of the present study approximated the educational attainment characteristics of their respective cohorts, with the exception of the youngest (20–24 years) and oldest (75+ years) groups, which had approximately 3 more years of education than expected.

The subjects were also asked to provide a subjective evaluation of their own health, vision, and hearing compared to other people their age. At least 90% of the subjects in all three age groups rated themselves as moderately good, good, or very good on these characteristics.

Ability Measures

The ability measures consisted of a battery of 12 tests selected to represent six primary mental ability factors: Induction, Memory Span, Associative Memory, Associational Fluency, Ideational Fluency, and Verbal Comprehension (Ekstrom, French, Harman, & Dermen, 1976). The factors, representing several aspects of verbal intelligence and memory, were chosen on the basis of their potential relevance to memory performance (Horn & Donaldson, 1980; Hultsch, Nesselroade, & Plemons, 1976). Two specific tests representative of each primary mental ability were selected from published batteries, yielding a battery of 12 tests, shown in Table 1. In some instances, the format of the tests was modified slightly in order to clarify the instructions and simplify the response modes for older adult subjects. None of the modifications was considered extensive enough to affect the measurement validity of the tests.

Text Materials

The text materials consisted of four narratives, each approximately 500 words in length. The narratives were

Table 1
Intellectual Ability Measurement Battery

Primary mental ability	Representative marker test	Source
Induction	Letter Sets Test[a]	Ekstrom, French, Harman, & Dermen (1976)
Induction	Letter Series Test	Thurstone (1962)
Memory Span	Visual Number Span[a]	Ekstrom et al. (1976)
Memory Span	Auditory Number Span Backwards[a]	After Ekstrom et al. (1976)
Associative Memory	Object Number Test[a]	Ekstrom et al. (1976)
Associative Memory	Memory for Words II	Kelley (1964)
Associational Fluency	Controlled Associations Test[a]	Ekstrom et al. (1976)
Associational Fluency	Figures of Speech Test[a]	Ekstrom et al. (1976)
Ideational Fluency	Topics Test[a]	Ekstrom et al. (1976)
Ideational Fluency	Theme Test[a]	Ekstrom et al. (1976)
Verbal Comprehension	Vocabulary I[a]	Ekstrom et al. (1976)
Verbal Comprehension	Advanced Vocabulary[a]	Ekstrom et al. (1976)

[a] Part I only.

abstracted from magazine articles, and each dealt with a life event experienced by a central female character. The events included bearing a first child, recovering from an injury sustained in an automobile accident, returning to school and beginning a new career, and coping with a family financial problem.

Kintsch's (1974) system was used to represent the meaning of the texts. Within this system, the meaning of a text is represented by a structured set of propositions known as a text base. A proposition consists of a predicate and one or more arguments. Predicates tend to be verb forms and specify a relation among the arguments. Arguments are word concepts or other propositions themselves. A propositional analysis of each text was done according to the criteria developed by Kintsch (1974) and elaborated by Turner and Green (1978). Each of the texts contained from 221 to 248 propositions.

Procedures

The text recall and ability tasks were administered to small groups of 3 to 10 individuals over three occasions. During the first session, the subjects were asked to read and remember the texts and to complete four of the ability measures. The texts were presented in typewritten booklets. The order of the stories was partially counterbalanced, with each text occurring once in each ordinal position. Prestige Pica 10-pitch type was used in order to minimize possible sensory difficulties. The subjects were instructed to read each of the four texts at their own pace. Recall was tested after each text with subjects writing their recall on lined pages in the booklet. It was emphasized to the subjects that verbatim recall was not required. Following the text recall task, all subjects completed the Vocabulary I, Controlled Associations, Letter Sets, and Advanced Vocabulary tests. The tests were administered in invariant order and under the time limits specified in the original source.

One week following the original session, the subjects were asked to remember the texts again and to complete the remaining eight ability measures. For the second recall test, the title of each narrative was printed on a lined page of the recall booklet, and the subjects were instructed to write their recall on the page. Again, it was emphasized to the subjects that verbatim recall was not required. Following the text recall task, all subjects completed the Theme, Letter Series, Memory for Words, Visual Number Span Forward, Figures of Speech, Object Number, and Auditory Number Span Backwards tests. The tests were administered in invariant order and under the time limits specified in the original source.

Finally, 4 weeks following the original session, the subjects were asked to remember the texts a third time. The procedures followed for this final recall test were the same as those used for the second recall test.

Recall Protocol Scoring

Each recall protocol was checked against the propositions of the original text base in order to determine whether each proposition was expressed in the protocol. In the scoring system used, a proposition was scored as correctly recalled if it contained the "gist" of the proposition's meaning (Turner & Green, 1978). Thus, overspecified or generalized relations and arguments were scored as correct (if substantively correct). If the subject made an error in one proposition and then repeated the error in a subordinate proposition, the subordinate proposition was scored as correct to avoid counting errors more than once.

A separate study was conducted to determine the interrater reliability of this scoring system. Twelve protocols were randomly selected for each of the four stories and independently scored by two scorers. With an average of 230 propositions per story, there were approximately 2,760 scoring decisions made by each scorer. There was 95.9% agreement between the two scorers on whether a proposition should or should not be scored as correctly recalled. During the course of the scoring, one of the scorers had to be replaced. Accordingly, interrater reliability was assessed a second time for the new pair of scorers using the same procedures as before. The analysis revealed 93.8% agreement between the two scorers.

LISREL Methodology

The analyses reported in this paper are based on the factor analysis measurement model in LISREL (Joreskog & Sorbom, 1979). In modified LISREL notation, the measurement model expresses the covariance matrix of the observed variables in the populations, Σ, as

$$\Sigma = \Lambda\Psi\Lambda + \Theta, \quad (1)$$

where Λ is a $p \times m$ matrix of factor loadings, Ψ is the covariance matrix of the factors, and Θ is the covariance matrix of the residual (unique) components.

It is necessary to specify a model that has a unique solution for the parameters by placing a sufficient number of restrictions on the equations in (1) to identify the remaining unknowns. Restrictions are specified by either fixing parameters to a known value a priori (e.g., requiring that a variable is unrelated to a factor by fixing its regression in Λ to 0), or constraining a set of two or more parameters to be equal. One of the advantages of LISREL's equality constraints is that parameters may be constrained equal between different age groups. Overidentified models provide a likelihood ratio χ^2 test statistic that may be used to test the goodness of fit of the model. Differences in χ^2 between two alternative models are particularly useful for hypothesis testing. For example, the difference in χ^2 between a model forcing all text memory correlations with intelligence factors to be zero, and a model freely estimating the correlations, is a likelihood ratio test of the null hypothesis that the correlations are in fact zero in the population.

In the present data analysis, the small sample sizes of 50 subjects per age group require some caution in the use of LISREL and χ^2 testing. First, the assumption that the sample covariance matrices provide asymptotic estimates of the population covariance matrices may be violated. The consequences of violating this assumption include the possibility that model parameters may be somewhat sample specific and may not be replicable in larger, independent samples. Thus, the analyses reported here should be considered exploratory attempts at model building, which must be replicated and extended on new samples. Second, the small sample sizes means that the likelihood ratio tests have relatively low statistical power. The greater possibility of Type II errors creates a special problem for LISREL models—it is possible to accept a set of restrictions that are in fact untenable in the population, and would be shown to be so had a larger sample size been employed. The analyses reported here were conducted with careful attention to this issue.

In multiple groups analysis, it is necessary to estimate factor models using covariance metric and sample covariance matrices rather than to analyze separately standardize correlation matrices. Standardization could obscure invariant relationships because of group differences in observed variances (see Cunningham, 1978; Jöreskog, 1971). The analyses reported here were all conducted in covariance metric, and LISREL's maximum likelihood parameter estimates and their standard errors are therefore in unstandardized form. Because standardized statistics are easier to interpret, we also report parameter estimates that have been rescaled to standardized metric.[1]

Results

The data analysis consisted of two parts. The first part examined differences in text recall as a function of age, delay interval, and story. The second part assessed relations between text recall performance and the ability variables.

Age Differences in Text Recall

In order to examine differences in gist recall, a 3 (age) × 3 (delay interval) × 4 (story) mixed-model analysis of variance (ANOVA) with repeated measures on the last two factors was performed on the percentage of correctly recalled propositions.[2] The analysis revealed significant main effects of age, $F(2, 147) = 33.89$, $p < .001$, and delay interval, $F(2, 294) = 323.38$, $p < .001$. Neuman-Keuls analyses conducted at $p < .05$ revealed that the younger adults ($M = 17.33$) recalled a significantly greater percentage of propositions than the older adults ($M = 6.20$). The two younger groups and the two older groups did not differ significantly. Neuman-Keuls analyses also revealed that the participants recalled a significantly greater percentage of propositions at immediate recall ($M = 18.49$) than following delays of 1 ($M = 10.34$) or 4 weeks ($M = 8.69$). There was no significant difference between the 1- or 4-week intervals.

The analysis also revealed a significant interaction of age with delay interval, $F(4, 294) = 4.27$, $p < .01$, shown in Figure 1. Neuman-Keuls analyses indicated that at all three recall tests, the younger and middle-aged adults recalled a significantly greater percentage of propositions than the older adults. The two younger groups did not differ significantly. As shown in Figure 1, however, the differences between the age groups are somewhat greater at the immediate recall test

[1] We do not report the full model specification or the maximum likelihood estimates for all models. Readers interested in a more detailed description of the specification and tables of maximum likelihood estimates for all models should write to C. Hertzog.

[2] Mixed-model F tests may be positively biased if the circularity assumption is violated; however, multivariate significance tests for repeated measures effects agreed with the mixed model tests in all cases.

than they are at the delayed recall tests. Nevertheless, this interaction may actually reflect the fact that the older adults are exhibiting a "floor effect" at the later delay intervals. Thus, given the similarity of the curves in Figure 1, the most reasonable interpretation is that the rate of forgetting is similar for all three age groups.

Finally, the analysis also revealed a significant main effect of story, $F(3, 441) = 47.94$, $p < .001$, and an interaction of this variable with age, $F(6, 441) = 3.53$, $p < .01$. These effects were a function of the fact that the younger and middle-aged adults recalled one of the stories better than the others. Because the four stories were not selected along any a priori dimensions, these effects were not interpreted further.

Intelligence-Text Recall Relationships. Age × Ability Interactions

Our approach to testing Age × Ability interactions in text recall performance involves (a) the use of factor analysis to define ability factors; (b) the computation of ability factor scores using the factor regression method; and (c) the joint regression analysis of text recall performance on ability factors, age, and Ability × Age interaction terms.

Analysis of intellectual variables. We first confirmed the expected pattern of age differences in psychometric intelligence by computing a multivariate analysis of variance (MANOVA) on the age factor for all 12 ability variables, using a subsample of 143 subjects with complete psychometric data. There were significant age differences in intelligence, approximate $F(24, 258) = 8.25$, $p < .001$. Univariate tests (not reported here in the interests of brevity) indicated significant age differences on all subtests except Vocabulary I, with the largest age differences on Letter Series and Letter Sets. Thus, there were significant age differences in ability (consistent in pattern and magnitude with previous reports in the literature; e.g., Horn & Donaldson, 1980),

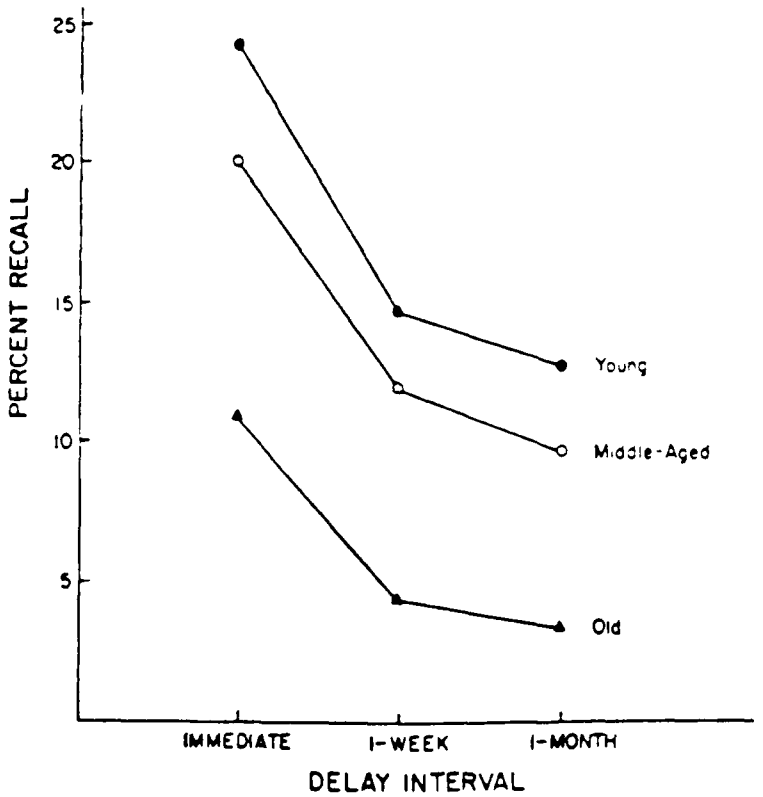


Figure 1 Percentage of propositions recalled as a function of age and delay interval averaged over stories.

which could contribute to the observed age differences in text memory performance.

In order to address the intelligence-text recall relationship, we conducted a confirmatory factor analysis on the 12 psychometric subtests. As indicated above, the intelligence subtests were originally selected in order to measure the six primary ability factors listed in Table 1. The initial confirmatory factor analysis specified this six-factor model with all loadings except those listed in Table 1 fixed to zero. The results indicated that the primary ability factor model was "overfit," with a small χ^2 and factor correlations uniformly high (generally in the .7 to .9 range). These results were problematic for any attempt to correlate text memory performance with primary ability factors in order to identify ability-specific differences in relations with text recall performance, because each subtest has a substantial regression on a second-order general intelligence factor (g). If age groups differ in the magnitude of relationship of primary ability factors to a second order g factor, we could detect differences in correlations between text recall and two primary ability factors (e.g., Verbal Comprehension vs. Memory Span) even when the only meaningful relationship was between g and text recall. We therefore opted for a factor analysis model that directly modeled g as one of the factors and then represented the other ability factors as residual or group factors. We consider this model to be a defensible representation of the factor structure that could be meaningfully used to determine whether intelligence-text recall relationships were a function of g or more specific factors such as Associative Memory, Verbal Comprehension, and Associational Fluency.

We proceeded to estimate a model specifying a general intelligence factor plus four specific factors: Verbal Comprehension, Verbal Productive Thinking, Memory Span, and Associative Memory.[3] Our initial results forced several modifications of this model. Although the Memory Span and Associative Memory variables have been conceptualized as loading on the Horn's Secondary Acquisition and Retrieval factor (Horn & Donaldson, 1980), we did not find a Memory Span factor independent of g. The results also showed that the Theme subtest did not load on the Verbal Productive Thinking factor. Subsequent models fixed this factor loading to zero. The fit of the modified four factor model was excellent, $\chi^2(44, N = 143) = 49.97$, $p = .25$, $F = .176$, indicating the model was a plausible representation of the factor structure in the entire sample.

We also examined the issue of age-group differences in factor structure. If different factor models were required to account for the covariances among the psychometric measures, the measurement equivalence of these ability measures across age groups would be called into question. An important implication of a lack of factorial invariance for the present analysis would be that the relationship between ability factor scores and text recall performance could not be examined by regression analysis on the entire sample, because the relationship of measures to ability factors would vary with age.

As shown by Meredith (1964), group selection from a population for which a common factor model holds will yield an invariant raw score (unstandardized) factor pattern matrix, but unique variances, factor variances, and factor covariances may vary because of selection effects. An implication of Meredith's work is that empirical evidence indicating an invariant raw score factor pattern matrix is consistent with a simple selection model, which, if true, would justify further analysis of ability-text recall relationships based on the single group factor solution.

We therefore estimated a series of simultaneous three-group models specifying the same four factors: g, Verbal Productive Thinking, Verbal Comprehension, and Associative Memory, testing the hypotheses of between-group equivalence in Θ, Ψ, and Λ. The hypotheses of group equivalence in Ψ and Θ were not rejected, $\chi^2(14, N = 143) = 22.10$, $.05 < p < .10$, and $\chi^2(24, N = 143) = 32.14$, $p < .10$, respectively. However, the absolute χ^2 test was statistically significant

[3] Induction was not estimated as a group factor because of its close relationship to g (Vernon, 1979); Verbal Productive thinking is Horn's second order factor combining Associational and Ideational Fluency (Horn & Donaldson, 1980); we used Verbal Productive thinking because the estimated correlation of the two primary ability factors exceeded .9 in the six-factor model.

Table 2
Results of the Ability Factor Analysis in the Three Age Groups

Subtest	Factor Pattern Matrix (Λ)				Unique Variance Θ		
	g	VC	VPT	AM	Young	Middle	C
Vocabulary 1	.59	.65	0*	0*	.31	.19	
Advanced Vocabulary	.51	.75	0*	0*	.15	.19	
Controlled Associations	.43	0*	.62	0*	.63	.36	
Figures of Speech	.36	0*	.37	0*	.73	.67	
Topics	.27	0*	.38	0*	.87	.71	
Theme	.19	0*	0*	0*	.97	.95	
Forward Span	.47	0*	0*	0*	.80	.74	
Backward Span	.48	0*	0*	0*	.83	.70	
Object Number	.42	0*	0*	.37	.72	.62	
Memory for Words	.60	0*	0*	.70	0*	0*	
Letter Series	.87	0*	0*	0*	.25	.27	
Letter Sets	.65	0*	0*	0*	.59	.58	

Note. g = general intelligence factor; VC = Verbal Comprehension factor; VPT = Verbal Productive Thinking factor; AM = Associative Memory factor. All elements in Λ, and Θ, are rescaled to a quasistandardized correlation metric using the approach recommended by Jöreskog (1971).
* Fixed parameter.

for the model with all matrices invariant over groups, $\chi^2(200, N = 143) = 237.90$, $p < .05$. Given the small sample sizes, we elected to allow Θ and Ψ to vary freely over groups in subsequent models. A model allowing all factor loadings to vary freely over groups did not significantly improve the fit of the model.

Table 2 reports the scaled factor loadings, factor variances and covariances, and unique variances for the model requiring Λ to be invariant over groups but allowing group differences in Θ and Ψ. The model provided an adequate fit to the data, $\chi^2(164, N = 143) = 184.66$, $p > .10$. As can be seen from Table 2, *g* was marked by high loadings for the Induction subtests, Letter Sets and Letter Series, but there were significant loadings on all subtests. Relatively high loadings were found for the Verbal Comprehension subtests and for Memory for Words on *g* as well. The Verbal Comprehension factor was well defined by both subtests, whereas Verbal Productive Thinking was defined most predominately by Controlled Associations and Associative Memory was weighted toward Memory for Words.

Perhaps the most interesting results were found in the group differences in Ψ. Table 3 shows that, although the correlations between the Verbal Productive Thinking, Verbal Com-

Table 3
Results of Factor Analysis in the Three Age Groups: Factor Covariance Matrices (Ψ)

Factor	Young				Middle				Old			
	g	VC	VPT	AM	g	VC	VPT	AM	g	VC	VPT	AM
g	9.90	0*	0*	0*	16.86	0*	0*	0*	11.28	0*	0*	0*
VC	0*	4.10	.37	.50	0*	7.96	.68	.46	0*	10.29	.64	
VPT	0*	1.77	5.35	−.19	0*	6.97	12.96	.53	0*	5.53	7.16	
AM	0*	1.46	−.64	2.10	0*	1.73	2.56	1.81	0*	1.99	1.33	

Note. g = general intelligence factor; VC = Verbal Comprehension factor; VPT = Verbal Productive Thinking factor; AM = Associative Memory factor. All elements in Λ, and Θ, are rescaled to a quasistandardized correlation metric using the approach recommended by Jöreskog (1971). Correlations are above the diagonal.
* Fixed parameter.

prehension and Associative Memory factors were relatively modest in magnitude in the young group, they were generally larger in the middle-age and older groups.

For both the covariances and correlations, the largest differences seemed to be between the young group and the two older groups. In general, the middle-aged group was more variable in ability, although the old group was the most variable in Verbal Comprehension and the young group had the largest variance in Associative Memory. The tendency for the older group to have higher correlations among abilities than the middle-aged group is qualified by the group differences in variances: The covariances among abilities differ little between the two older groups.

Because the results from the multiple group analysis were consistent with the selection hypothesis discussed by Meredith (1964), pooling the data over age groups for further analysis was justified. We used LISREL's factor score regression matrix to estimate ability factor scores for the entire sample. Table 4 gives the calculated correlations among the factor score variables that agreed relatively well with the LISREL maximum likelihood estimates of the factor correlations. Note, however, that the substantial correlations between the specific ability factor scores, especially between Verbal Comprehension and Verbal Productive Thinking, create the possibility of suppression effects in the regression analysis.

Intelligence × Age interactions. The regression analysis of age and ability variables is equivalent to the analysis of covariance (ANCOVA) approach, but with the interaction between the independent variable (age) and the covariates (ability factor scores) explicitly represented in the design. Tests of such interaction terms are often treated only as tests of ANCOVA assumptions. However, the ANCOVA analogy is misleading here because the interactions provide the critical information regarding the consistency of age differences in text memory across levels of ability and are therefore of substantive interest in their own right (Cohen & Cohen, 1975).

The interaction variables were calculated by multiplication of two orthonormal contrasts across the age factor with the factor score variables after these latter variables had been transformed to z scores. The two contrasts selected compared (a) middle-aged with old subjects and (b) young subjects against the combined middle-aged with old age groups. The regression equations therefore included 14 independent variables organized in three sets: (a) the four ability factors, (b) two age contrasts, and (c) eight interaction variables representing the products of these first two sets of variables. A separate regression analysis was conducted for each of the three delay levels (immediate, 1 week, and 4 weeks). Before examining partial regression coefficients we calculated hierarchical significance tests of the increment R^2 for three sets of independent variables; the four ability factors, the two age contrasts, and the eight Ability × Age interaction variables.

The results of the hierarchical significance tests are given in Table 5. For each delay level, the overall R^2 was highly reliable, with the adjusted R^2 of greater than .5 for each equation. Thus, a large proportion of text recall variance was accounted for by the model. The increments to R^2 for the ability factors were large and significant for all three delay conditions, accounting for greater than 80% of the total R^2 in each case. The overall test of age differences was also significant at each delay condition, with R^2 smallest at immediate recall. Adjusted for shrinkage, age accounted for between 3% and 4% of the variance across all delay conditions. This is

Table 4
Factor Correlations for Intelligence Variables Single Group Analysis

Factor	*g*	VC	VPT	AM
g	1[a]	.054	.083	.094
VC	0	1	.739	.520
VPT	0	.598 (.115)[b]	1	.317
AM	0	.489 (.127)	.282 (.137)	1

Note *g* = general intelligence factor; VC = Verbal Comprehension factor; VPT = Verbal Productive Thinking factor; AM = Associative Memory factor. LISREL estimates and standard errors for the correlations among ability factors are given below the diagonal, and correlations among estimated factor scores are given above the diagonal.

[a] All zeroes and ones are fixed parameters in LISREL model.

[b] Standard errors are in parentheses.

Table 5
Summary of R^2 and Statistical Tests for Regression Analysis With Age and Intelligence Factors

Dependent variables (delay conditions)	Independent variables (Set)	R^2	$\tilde{R}^{2a}$	ΔR^{2b}	$\Delta \tilde{R}^{2c}$	F	df	p
Immediate	IQ	.452	.436	.452	.436	28.42	(4, 138)	<.00
	Age	.486	.463	.034	.027	4.46	(2, 136)	<.05
	IQ × Age	.557	.508	.071	.045	2.56	(8, 128)	<.05
	(Age alone)[d]	.255	.244	—	—	—	—	—
1 week	IQ	.478	.462	.478	.462	31.53	(4, 138)	<.00
	Age	.523	.502	.045	.040	6.43	(2, 136)	<.01
	IQ × Age	.570	.523	.047	.021	1.75	(8, 128)	> .05
	(Age alone)	.316	.306	—	—	—	—	—
4 weeks	IQ	.475	.460	.475	.460	31.21	(4, 138)	<.00
	Age	.516	.494	.041	.034	5.73	(2, 136)	<.01
	IQ × Age	.570	.523	.054	.029	2.01	(8, 128)	> .05
	(Age alone)	.322	.312	—	—	—	—	—

Note. IQ = intelligence factor scores. Regression statistics are for hierarchical regression entering three sets of independent variables: four intelligence factors, two age contrasts, and eight interaction variables.
[a] R^2 adjusted for shrinkage.
[b] Change in unadjusted R^2 from previous set.
[c] Change in shrinkage adjusted R^2 from previous set.
[d] R^2 for two age contrasts as only independent variables (i.e., ignoring intelligence).

clearly a major reduction in the prediction of test memory performance by age, because it accounted for between 20% and 30% of the variance when entered without the ability variables (see Table 5). Nevertheless, the analysis indicates there are age differences in text memory performance that are independent of intellectual ability.

The Ability × Age interactions were not consistently reliable, exceeding a 5% alpha level only for the immediate recall condition (although there were 10% level trends for both longer delay intervals). Thus, all three types of variables provided independent contributions to the total R^2, with the largest amount of variance accounted for by the ability-text memory relationships measured at the level of the ability factors; age differences in text memory covary highly with age differences in multiple intellectual abilities; however, age differences in text memory cannot be eliminated by partialing ability differences; and there may be Age × Ability interactions in text memory performance that qualify the existence of age main effects.

Table 6 reports the individual standardized regression parameter estimates and their standard errors, which may be used to calculate t tests of the null hypothesis that the regression weights are zero in the population. The pattern of results clearly differentiated the linear relationships of ability factor scores to text memory performance from the Age × Intelligence interaction effects. The general intellectual factor, *g*, provided the best independent prediction of text memory performance, but did not produce an interaction effect in conjunction with age at any delay level. The Verbal Comprehension and Associative Memory factors also provided independent prediction of text memory performance, although at a much smaller level of magnitude. Verbal Productive Thinking did not provide statistically reliable independent prediction of text memory performance. Of course, the relatively small independent contributions of the ability factors other than *g* are in part a function of mutual inhibition, considering the high intercorrelation between Verbal Comprehension and Verbal Productive Thinking.

In contrast to the simple linear ability effects, Verbal Productive Thinking contributed most to the significant overall interaction at immediate recall, interacting with both age contrasts. The direction of effects was in the predicted direction, with age differences between all three groups were smaller at higher

Table 6
Standardized Regression Parameters for Three Delay Conditions on Intelligence Factors and Age

Source	Delay condition (β)		
	Immediate	1 Week	4 Weeks
VC	.23 (.11)**	.26 (.11)**	.18 (.11)****
VPT	.05 (.10)	.00 (.09)	.10 (.10)
AM	.14 (.07)**	.13 (.07)*	.13 (.07)*
g	.39 (.09)****	.38 (.09)****	.37 (.09)****
AGE1	.18 (.09)*	.27 (.09)***	.24 (.09)***
AGE2	.24 (.11)**	.34 (.11)***	.37 (.11)****
VC × AGE1	.16 (.11)	.08 (.11)	.07 (.11)
VPT × AGE1	−.29 (.10)***	−.10 (.10)	−.15 (.10)
AM × AGE1	.02 (.07)	.04 (.07)	.10 (.07)
g × AGE1	.08 (.08)	.13 (.08)	.12 (.08)
VC × AGE2	.29 (.10)***	.28 (.10)***	.22 (.10)**
VPT × AGE2	−.25 (.09)***	−.13 (.09)	−.03 (.09)
AM × AGE2	−.16 (.07)**	−.11 (.07)	−.15 (.07)**
g × AGE2	.02 (.08)	0.0 (.08)	−.01 (.08)
AGE 1 (alone)	.37 (.07)****	.40 (.07)****	.43 (.07)****
AGE 2 (alone)	.35 (.07)****	.39 (.07)****	.37 (.07)****

Note VC = Verbal Comprehension, VPT = Verbal Productive Thinking, AM = Associative Memory, *g* = General Intelligence, AGE1 = first age contrast (middle aged vs. old), AGE2 = second age contrast (young vs. middle aged/old).

[a] Standard errors are in parentheses.

Significance levels for *t* test of H_0: $\beta = 0$ are denoted as follows: * $p < .10$. ** $p < .05$. *** $p < .01$. **** $p < .001$.

levels of fluency. The other significant interaction terms involved only the contrast between the young group and the two older groups for both Verbal Comprehension and Associative Memory. The pattern of interaction found for Associative Memory was similar to Verbal Productive Thinking. However, to our surprise we discovered that the pattern was actually reversed for Verbal Comprehension—age differences appeared to be greater at the higher levels of verbal ability! We verified the direction of this relationship through examination of a bivariate scatterplot and a regression using the original vocabulary subtests. This further analysis indicated that the strength of the effect was in part a function of classic suppression; the high positive correlation between Verbal Comprehension and Verbal Productive Thinking, combined with the opposite directions of interactions of each of the two ability factors with age, helped produce the statistically reliable positive regression coefficient for the Age × Verbal Comprehension interaction. Nevertheless, the direction of the effect, even in the bivariate plots, definitely showed increasing age differences at the highest levels of Verbal Comprehension.

This Age × Verbal Comprehension interaction term was statistically reliable at all delay levels. In contrast, the Age × Verbal Productive Thinking and Age × Associative Memory interactions were not significant at the longer delay intervals. Indeed, considering all the interaction terms together, it is clear that the interaction effects are at best small in magnitude and should not be given great interpretive weight.[4] However, when one considers that the reliable interaction effect for Verbal Comprehension was the reverse of the predicted relationship, it appears safe to conclude that the hypothesis of reduced age differences at higher ability levels was not supported by the data.

[4] In fact, it is possible to reduce the size of the interaction term for Verbal Productive Thinking by changing the factor model specification. The interaction is most strongly present (at the level of single subtests) for Controlled Associates; rotating the factor towards the Ideational Fluency subtests pushes the Age × Verbal Productive Thinking interactions below significance. Because the model used in this analysis is more parsimonious as a representation of the specific verbal factors, we have reported its results alone. Nevertheless, the evidence for interaction effects indicates that any effects in this population are relatively small in magnitude.

Intelligence-Text Recall Relationships: Correlational Analyses

The correlational analyses using LISREL were designed to explicate the relationship between text recall performance and psychometric intelligence in the three age groups. Our interest was in determining whether group differences in correlations among subtest scores and text memory (not reported here) reflected differential relationships of text recall with underlying dimensions of intelligence for the three age groups.

In order to examine the text recall correlations with intelligence, we introduced Text Recall as an additional factor in the factor model of the intelligence subtests. This model allows us to represent the covariances between the text recall variables and intelligence subtests as being mediated through the covariances between the text recall and intelligence factors, which were modeled in Ψ. We tested the ability/text recall relationships with the immediate recall data. The results were then replicated at the two longer delay intervals.

A first model forcing all four convariances between Text Recall and the four intelligence factors to equal zero provided a poor fit to the data, $\chi^2(321, N = 143) = 469.19$, $p < .001$. An alternative model allowing the covariances to be freely estimated fit considerably better, $\chi^2(309, N = 143) = 386.69$, $p < .01$. The difference in χ^2 tested the (multivariate) null hypothesis of zero correlations between Text Recall and the intelligence factors. This hypothesis was rejected, $\chi^2 = 82.50$ (12, $N = 143$) $p < .001$. We also tested the null hypothesis of group equivalence in the text recall–intelligence correlations by introducing a scaling vector in the model (thus allowing for group differences in variances) and constraining the scaled Text Recall ability covariances to be equal for the three age groups. This model produced a significant increase in χ^2 (20.99 with 8 *df*, $N = 143$ $p < .01$). The multivariate null hypothesis of equal correlations between age groups was therefore rejected.

Table 7 reports the Ψ matrices for the three groups, including the rescaled correlations between Text Recall and the four intelligence factors. The group differences in the text recall–intelligence correlations form an interesting pattern. In the young and middle aged groups, there is a statistically reliable correlation between *g* and Text Recall (.58 and .52, respectively). There is also a statistically reliable correlation between Verbal Comprehension and Text Recall in the young group ($r = .38$). However, this correlation was only .22 in the middle-aged group, less than the .31 correlation between Text Recall and Associative Memory. The correlational pattern in the old group is completely divergent. For the old adults, the correlation between *g* and Text Recall was not statistically reliable, but the remaining correlations between Text Recall and the other intelligence factors were statistically significant. Indeed, the correlation between Verbal Productive Thinking and Text Recall was .86, which was unexpectedly high. Given the definition of the other intelligence factors as being orthogonal to *g*, the results in the old group indicate that, in spite of the higher magnitude of the simple correlations among all the intelligence subtests, Text Recall performance was more highly correlated with the specific factors related to verbal intelligence and memory than to general intelligence. This was not the case in the young and middle-aged groups. As can be seen in Table 7, the pattern of differential correlations between age groups replicated at the longer delay intervals.

We also assessed the hypothesis that the lower levels and greater variability of years of education in the old group produced the differences in text recall–intelligence correlations. This was accomplished by partialing years of education from the factor correlations and examining the residual correlations. These residual correlations were highly similar to the original correlations, ruling out group differences in years of education as the determinant of age differences in text recall–intelligence correlations.

Discussion

The present data indicate that there are substantial age-related differences in the amount of information recalled from meaningful texts. These results are consistent with those of other studies that have examined the text recall performance of adults with relatively modest levels of education (Cohen

1979; Dixon et al., 1982; Zelinski et al., 1980). The present data also indicate that there is little evidence for age-related differences in the rate at which information about meaningful text is forgotten. These results are also consistent with most previous studies (e.g., Dixon et al., 1982; Gordon & Clark, 1974). Within this context, however, the present study suggests some important conclusions about the role of intellectual ability factors in age-related differences in text recall performance.

It has been repeatedly suggested that adult age differences in text performance may depend on the subjects' level of verbal ability because several studies have demonstrated

Table 7
Text Memory/Intelligence Correlations for Within Age Groups for Three Delay Conditions

Age	Parameter	*g*		Intelligence factors: VC		VPT		AM	
				Immediate					
Young	σ_{IQ}^2[a]	8.91		3.68		4.93		0.94	
	σ_{TM}[b]	36.31**	(11.87)	15.36*	(7.65)	−14.10	(11.0)	3.19	(3.18)
	r_{TM}[c]	.58		.38		−.30		.15	
Middle aged	σ_{IQ}	16.19		7.33		12.00		0.79	
	σ_{TM}	54.69**	(19.87)	16.09	(12.33)	12.25	(16.79)	7.30	(4.37)
	r_{TM}	.52		.22		.13		.31	
Old	σ_{IQ}	10.56		9.69		6.81		.91	
	σ_{TM}	9.84	(9.04)	33.93**	(11.98)	43.29***	(12.12)	9.55*	(4.19)
	r_{TM}	.16		.57		.86		.52	
				1 week					
Young	σ_{IQ}^2	9.99		4.03		5.07		1.01	
	σ_{TM}	21.41**	(7.96)	14.65**	(5.61)	−2.21	(7.53)	3.33	(2.36)
	r_{TM}	.47		.50		−.07		.23	
Middle aged	σ_{IQ}^2	17.35		7.29		12.35		0.85	
	σ_{TM}	39.70**	(13.24)	8.61	(7.39)	9.28	(10.42)	4.30	(2.69)
	r_{TM}	.56		.19		.16		.28	
Old	σ_{IQ}^2	11.31		9.63		7.15		0.96	
	σ_{TM}	6.16	(4.00)	9.92**	(4.28)	9.06*	(4.03)	3.59*	(1.59)
	r_{TM}	.26		.45		.48		.51	
				4 weeks					
Young	σ_{IQ}^2	10.40		4.22		4.54		1.00	
	σ_{TM}	16.05*	(7.07)	11.62*	(4.94)	8.46	(6.46)	1.24	(1.94)
	r_{TM}	.39		.45		.32		.10	
Middle aged	σ_{IQ}^2	17.64		7.56		9.67		0.84	
	σ_{TM}	33.91**	(11.28)	6.90	(6.30)	4.63	(8.02)	4.52	(2.41)
	r_{TM}	.57		.18		.11		.35	
Old	σ_{IQ}^2	11.41		10.05		5.30		0.95	
	σ_{TM}	4.30	(3.00)	8.75**	(3.49)	10.61**	(3.39)	2.48*	(1.17)
	r_{TM}	.23		.49		.82		.45	

Note. *g* = General Intelligence; VC = Verbal Comprehension; VPT = Verbal Productive Thinking; AM = Associative Memory.
[a] Variance of intelligence factor.
[b] Covariance of intelligence with text memory (standard error in parentheses).
[c] Correlation of intelligence with text memory.
Significance levels for H_0: $\sigma_{TM} = 0$ denoted as follows: * $p < .05$. ** $p < .01$. *** $p < .001$.

that age differences are present when subjects of low to medium verbal ability are examined and absent when subjects of high verbal ability are examined (Dixon, et al., in press; Meyer & Rice, 1983; Taub, 1979). However, the results of the present analysis suggest that the potential contribution of ability factors to age-related differences in text recall performance is more complex than previous reports might indicate.

First, it is apparent that abilities other than Verbal Comprehension are predictive of text recall performance. In particular, the present results suggest that general intelligence, Verbal Productive Thinking, and Associative Memory also correlate with individual differences in text recall performance. In fact, the ability with the largest overall relationship with text memory performance in the single group analysis was *g*, not Verbal Comprehension. Second, the present results show that age differences in text memory performance covary highly with age differences in intellectual abilities. The regression analyses indicated that age differences in text memory performance are drastically reduced, but not eliminated, when partialed for intellectual ability. Third, and perhaps most significantly, the present results do not support the notion that there is an Age × Verbal Comprehension interaction across the range of verbal abilities such that age differences are progressively reduced with higher ability level, as might be suggested by the results cited above. If anything, we found evidence for larger age differences at the highest Verbal Comprehension levels present in our sample. The type of interaction predicted by the previous work with extreme groups designs was only found in the immediate recall condition for Verbal Productive Thinking and Associative Memory; moreover, the small magnitude of the interaction effects and the transience of the relationship with respect to delay interval suggests, at minimum, that such interactions should be interpreted conservatively.

The present results need not be viewed as contradictory to previous findings if we allow for the fact that the population studied here is a community population that apparently contains small proportions of the extreme high ability/highly educated elderly. It may well be the case that age differences are smaller only at the highest ability or educational levels, or alternatively, that there is a small, relatively intact subpopulation of able elderly who show little decline in text memory performance. Comparisons of such a select subpopulation with young adult groups might well yield little age differences. Nevertheless, the present results speak to the issue of the generality of the results from the previous extreme groups comparisons. For the ability ranges studied here, the interaction effects do not suggest the elimination of the age differences at higher ability levels.

The final complexity in ability–text memory relationship discovered in the present study is the shift in patterns of within-group correlations between text memory and intellectual ability factors across the three age groups. In the case of the young and middle-aged adults, the largest correlations of text memory and ability factors occurred with *g* and Verbal Comprehension. However, in the case of the old adults, the largest correlations involved Verbal Productive Thinking, Verbal Comprehension, and Associative Memory. General intelligence is of little value in predicting text recall performance in the elderly. Thus, with increasing age, text recall performance is increasingly related to specific intellectual abilities including Verbal Productive Thinking and Associative Memory as well as Verbal Comprehension.

The reduced correlation between *g* and text memory performance in the old group is rather surprising. One of the consistently replicated findings in the literature on adult age differences in the factor structure of psychometric intelligence is that older populations have a less-differentiated factor structure than younger populations, usually manifested in a higher correlation among primary ability factors (e.g., Baltes, Cornelius, Spiro, Nesselroade, & Willis, 1980; Cunningham, 1980). A developmental hypothesis that has derived from this pattern is that of reintegration or de-differentiation of intelligence with aging, such that individual differences in cognitive activity are determined less by specific skills (as represented by the range of primary intellectual abilities) and more by general cognitive efficiency (Reinert, 1970). Some researchers (e.g., Birren, Woods, & Williams, 1979) have drawn a parallel to other studies suggesting a general slowing of cognitive speed with aging and have interpreted de-differen-

tiation of intelligence as an indication of the predominant importance of central nervous system integrity in determining individual differences in older populations. If this interpretation were taken to its logical extreme, we would predict that general intelligence should have a *higher* correlation with text recall performance in the elderly than in any other population, yet the pattern of effects in this study is in the opposite direction. Apparently not all forms of cognitive activity increase their correlation with *g* over the adult life span.

We should note, however, that text recall did correlate significantly with the other three intelligence factors (which in turn were highly intercorrelated). This pattern of effects might be taken to indicate that a second-order verbal intelligence factor, uncorrelated with *g*, correlates with text recall in the old group. This shift in correlations is provocative, but some caution is in order given the relatively small sample sizes. Certainly replication of these differences in larger samples would be a necessary part of any attempt to extend and explain these findings. We note, however, that Hultsch et al. (1976) found higher correlations between psychometric tests of Associative Memory and learning performance in an older sample than in a young group of subjects. Although the experimental tasks were not particularly comparable between the two studies, the similar shift in correlations lends additional validity to the present results.

We are inclined to view the shift in correlational pattern as a developmental phenomenon meriting further study. However, one could also argue that the group differences might have been produced artifactually by differential selection. For example, group differences in text recall–intelligence correlations could be a function of group differences in variables such as education. We found no indication that educational differences account for the shift in correlational patterns, but we cannot rule out other types of selection effects.

Assuming that the increased correlations actually do reflect some type of developmental phenomenon, how might it be characterized? Of the several possibilities, let us mention two. The first is that the results may be a function of age-related differences in strategies used to process the texts. There is recent evidence for such differences. For example, Rice and Meyer (1983) found younger and middle-aged adults are more likely than older adults to use a strategy that emphasized serial retrieval of information based on an understanding of the paragraph structure of the text. In contrast, older adults were more likely than younger and middle-aged adults to rely on a simpler strategy that emphasized the identification of the main ideas of the text. To the extent that different intellectual abilities support such different strategies, a changing pattern of correlations as a function of age would be produced. In this instance, then, abilities are functioning as indirect markers of strategy use.

A second explanation of the shift in correlational patterns involves the concept of differential loss of abilities that relate to text memory performance. From this perspective, most young persons would have sufficient semantic processing skills and memory for words to perform adequately on text comprehension and recall tasks. Thus, individual differences in text memory performance would not be predicted by individual differences in intellectual abilities. In older populations, on the other hand, it is possible that a subgroup of older persons would have suffered a sufficient level of decline in their semantic processing skills to cause declines in text recall performance, whereas other older persons would have maintained their skills. Such a pattern would increase the predictive value of individual differences in associative memory and other semantic processing skills for text recall performance in the older groups because the range of individual differences in semantic processing skills would include levels that would have an adverse impact on performance on text recall tasks.

This interpretation is consistent with findings from the psychometric literature concerning the terminal decline phenomenon (Riegel & Riegel, 1972). It is well known that, on average, older persons are more likely to decline in primary abilities related to fluid intelligence, spatial visualization, or perceptual speed, but are likely to maintain levels of crystallized intelligence, including numerical and verbal abilities (see Horn & Donaldson, 1980). However, the literature on nonnormative pathological decline prior to death

shows that the decline does not spare verbal skills. Indeed, the phenomenon of terminal decline is best identified by the fact that vocabulary and knowledge-oriented tests, which normally remain relatively stable, decline (e.g., Blum & Jarvik, 1974). From the differential loss perspective, one would argue that declines in text recall performance are relatively nonnormative, in the sense that they cannot be expected for all (or perhaps even a majority of) elderly individuals. Instead *only some* individuals in the older population exhibit a sufficiently large decline in semantic processing skills to adversely affect text recall performance. Such a phenomenon could account for (a) the shift in the correlational pattern of intelligence and text recall over different age groups; (b) the inconsistency in the literature of studies finding age differences in text memory performance, because finding mean differences would depend on the relative proportion of the declining elderly subpopulation sampled; and (c) the differential probability of finding age differences in text memory among groups partitioned by high and low verbal ability.

Finally, some combination of these explanations is possible. For example, differential decline may be the source of age-related differences in encoding or retrieval strategies. Such a possibility is consistent with recent findings reported by Spilich (1983). He found evidence of poorer text performance in "normal" elderly compared to younger adults, but not qualitative age differences in text processing strategies. In contrast, he found evidence for such qualitative differences between the "normal" elderly and memory-impaired elderly.

Thus, poor text recall performance in later life may reflect two different phenomena that are hopelessly confounded in a cross-sectional design: First, low-ability subjects whose poor text performance reflects the continuation of poor verbal skills over the life span, and second, low-ability subjects whose poor text performance reflects a loss of verbal skills from previously higher levels. Clearly, a short-term longitudinal study examining changes in intellectual abilities and text recall performance in middle-aged and elderly adults would be required to examine these possibilities.

In summary, the present study clearly suggests that (a) text recall performance in adulthood is predicted not only by Verbal Comprehension, but by multiple abilities; (b) that age differences in text memory performance overlap highly with age differences in multiple intellectual abilities, although ability differences do not fully account for the age differences in text recall; (c) modest Age × Intellectual ability interactions may exist, but the pattern of Age × Ability interactions does not suggest decreasing age differences in text recall with increasing ability across the range of the ability distribution; and (d) that there are differences in the pattern of within-age-group intelligence–text recall performance correlations. The results may well be problematic for a representation of text recall performance declines as simply quantitative changes in an otherwise qualitatively invariant cognitive process. They also suggest that cognitive psychologists should carefully examine the semantic processing factors associated with text recall performance, keeping in mind that accounting for individual differences in decline functions may be the critical feature needed to solve the problem.

References

Baltes, P. B., Cornelius, S. W., Spiro, A. III, Nesselroade, J. R., & Willis, S. L. (1980). Integration vs. differentiation of fluid-crystallized intelligence in old age. *Developmental Psychology, 16*, 625–635.

Birren, J. E., Woods, A. M., & Williams, M. V. (1979). Speed of behavior as an indicator of age changes and the integrity of the nervous system. In F. Baumeister (Ed.), *Bayer symposium IX: Brain function in old age* (pp. 10–44). Hamburg: Springer.

Blum, J. E., & Jarvik, L. F. (1974). Intellectual performance of octogenerians as a function of education and initial ability. *Human Development, 17*, 364–375.

Cohen, G. (1979). Language comprehension in old age. *Cognitive Psychology, 11*, 412–429.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Craik, F. I. M. (1977). Age differences in human memory. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 384–420). New York: Van Nostrand Reinhold.

Cunningham, W. R. (1978). Principles for identifying structural differences: Some methodological issues related to comparative factor analysis. *Journal of Gerontology, 33*, 82–86.

Cunningham, W. R. (1980). Age comparative factor analysis of ability variables in adulthood and old age. *Intelligence, 4*, 133–149.

Dixon, R. A., Hultsch, D. F., Simon, E. W., & von Eye, A. (in press). Verbal ability and text structure effects on adult age differences in text recall. *Journal of Verbal Learning and Verbal Behavior*.

Dixon, R. A., Simon, E. W., Nowak, C. A., & Hultsch, D. F. (1982). Text recall in adulthood as a function of level of information, input modality, and delay interval. *Journal of Gerontology, 37,* 358–364.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Gordon, S. K., & Clark, W. C. (1974). Application of signal detection theory to prose recall and recognition in elderly and young adults. *Journal of Gerontology, 29,* 64–72.

Harker, J. O., Hartley, J. T., & Walsh, D. A. (1982). Understanding discourse—a life-span approach. In B. A. Huston (Ed.), *Advances in reading/language research* (Vol. 1, pp. 155–202). Greenwich, CN: JAI Press.

Horn, J. L., & Donaldson, G. (1980). Cognitive development II: Adulthood development of human abilities. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development. A volume of review essays* (pp. 445–529). Cambridge, MA: Harvard University Press.

Hultsch, D. F., & Dixon, R. A. (1984). Text processing in adulthood. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 6, pp. 77–108). New York: Academic Press.

Hultsch, D. F., Nesselroade, J. R., & Plemons, J. K. (1976). Learning-ability relations in adulthood. *Human Development, 19,* 234–247.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt.

Kelley, H. P. (1964). Memory abilities: A factor analysis. *Psychometric Monographs,* No. 11.

Kintsch, W. (1974). *The representation of meaning in memory* Hillsdale, NJ: Erlbaum.

Krauss, I. K. (1980). Between- and within-group comparisons in aging research. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 542–551). Washington, D.C.: American Psychological Association.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Meyer, B. J. F., & Rice, G. E. (1981). Information recalled from prose by young, middle and old adults. *Experimental Aging Research, 7,* 253–268.

Meyer, B. J. F., & Rice, G. E. (1983). Learning and memory from text across the adult life span. In J. Fine & R. O. Freedle (Eds.), *Developmental studies in discourse* (pp. 291–306). Norwood, NJ: Albex.

Poon, L. W., Fozard, J. L., Cermak, L. S., Arenberg, D., & Thompson, L. W. (Eds.). (1980). *New directions in memory and aging: Proceedings of the George A. Talland memorial conference.* Hillsdale, NJ: Erlbaum.

Reinert, G. (1970). Comparative factor analytic studies of intelligence throughout the human life span. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology: Research and theory* (pp. 476–484). New York: Academic Press.

Rice, G. E., & Meyer, B. J. F. (1983, August). *Prose recall: Effects of aging, verbal ability, and reading behavior.* Paper presented at the meeting of the American Psychological Association, Anaheim, CA.

Riegel, K. F., & Riegel, R. M. (1972). Development, drop, and death. *Developmental Psychology, 6,* 306–319.

Simon, E. W., Dixon, R. A., Nowak, C. A., & Hultsch, D. F. (1982). Orienting task effects on text recall in adulthood. *Journal of Gerontology: 37,* 575–580.

Spilich, G. J. (1983). Life-span components of text processing: Structural and procedural changes. *Journal of Verbal Learning and Verbal Behavior, 22,* 231–244.

Taub, H. A. (1975). Mode of presentation, age, and short-term memory. *Journal of Gerontology, 30,* 56–59.

Taub, H. A. (1976). Method of presentation of meaningful prose to young and old adults. *Experimental Aging Research, 2,* 469–474.

Taub, H. A. (1979). Comprehension and memory of prose materials by young and old adults. *Experimental Aging Research, 5,* 3–13.

Taub, H. A., & Kline, G. E. (1978). Recall of prose as a function of age and input modality. *Journal of Gerontology, 33,* 725–730.

Thurstone, T. G. (1962). *Primary mental abilities. Grades 9-12, 1962 revision.* Chicago: Science Research Associates.

Turner, A., & Greene, E. (1978). The construction and use of a propositional text base. *JSAS Catalogue of Selected Documents in Psychology, 8,* 58. (Abstract)

U.S. Bureau of the Census. (1977). *Educational attainment in the United States: March 1977 and 1976.* Current Population Reports, Series P-20, No. 314.

Vernon, P. E. (1979). *Intelligence: Heredity and environment.* San Francisco: W. H. Freeman.

Zelinski, E. M., Gilewski, M. J., & Thompson, L. W. (1980). Do laboratory tests relate to self-assessment of memory ability in the young and old? In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging: Proceedings of the George A. Talland memorial conference* (pp. 519–544). Hillsdale, NJ: Erlbaum.

Received November 30, 1982
Revision received October 18, 1983 ■

DISTRIBUTION LIST

Office of Naval Research (Code 222)
Defense Technical Information Center (DTIC) (2)